Scan and Share 1.07-st
Tutorial on making e-books

written by V. and A.

2010

Contents

1 Introduction **3**
1.1 In brief . . . 3
1.2 Why make a scanned book, is OCR not good? . . . 3
1.3 How to get good quality of scans . . . 3

2 Scanning a book **5**
2.1 Setting up IrfanView for scanning . . . 7
2.2 Setting up VueScan for scanning . . . 9
2.3 Handwork while scanning . . . 9

3 Processing scans with ScanKromsator **13**
3.1 Draft run . . . 14
3.2 Set options . . . 15
3.3 Main run . . . 18
3.4 Processing color figures and photos . . . 18

4 Processing scans with ScanTailor **19**
4.1 Importing scan into ScanTailor . . . 20
4.2 Draft run . . . 21
4.3 More about processing steps . . . 22
4.4 Correct errors after the draft run . . . 23
4.4.1 Adjusting the content rectangle . . . 25
4.4.2 Adjusting the page alignment . . . 25
4.4.3 Adjusting the page sizes . . . 26

4.4.4 Adjusting the splitting 26
4.4.5 Adjusting the deskewing 28
4.4.6 Replacing scans in the project 28
4.5 Final run and final check-up 29
4.6 Working with picture zones 31

5 Encoding scans into DJVU **33**

6 Creating text layer with OCR **36**

7 Adding book covers and color plates **38**

8 Adding hyperlinks and bookmarks **38**

A Where to download software **40**

This document can be distributed for free. It is an expanded version of the "Scan and Share 1.07" tutorial. This tutorial now covers the new program, ScanTailor, as well as ScanKromsator. Some screenshots are in Russian because the software does not have any other localization. Screenshots of VueScan options are included now.

1 Introduction

This is a mini-tutorial about scanning books and making high-quality files out of them. This tutorial is intended for people who would like to make good-quality electronic books but do not know where to start. There are many ways to get good results by scanning; this text shows you some reasonably easy ways. The tutorial has step-by-step screenshots and assumes some familiarity with Windows. You may need to download and install a few programs (see Appendix A).

1.1 In brief

For the impatient reader: The process consists roughly of the following stages:

1. Scan every page in 300dpi greyscale, save to TIF. Save a backup of your scans!
2. Import images into ScanKromsator or ScanTailor, process images. Save a backup of the processed images at this stage!
3. Create a DJVU file out of processed images.
4. Add OCR and/or bookmarks to the DJVU file.

(It is most important to master the stages 1 and 2, since the processed images after stage 2 are much smaller than the initial scans, and you can send them to somebody else if you have trouble with stages 3 and 4.)

1.2 Why make a scanned book, is OCR not good?

Here I will be mostly talking about scanning of old books on science, mathematics, or technical books. For these books, OCR is not practical because these books contain too many equations, diagrams, graphs etc. No OCR program can accurately recognize this kind of material. The only solution is to scan and make images of all pages.

1.3 How to get good quality of scans

Such books are almost always printed purely in black/white, with perhaps very few pages having greyscale or color illustrations. For that kind of books, the highest quality of scanned e-books is achieved if one uses **600dpi** black/white images for most pages.[1] So you need to scan either directly in 600dpi black/white, or at **300dpi greyscale** and then process the scans to make them into 600dpi

[1] If you don't know what 600dpi means: it's called the *resolution* of the image and means the number of image points (pixels) per inch (dpi=dots per inch).

black/white.[2] If the book has a few pages with color illustrations, you will need to scan them separately in 300dpi 24-bit color mode. The same applies to colorful book covers that you also may want to scan.

Please note:

- **Never** scan at 300dpi black/white! The quality of the results is never as good as what you can get by scanning in 300dpi greyscale and following this tutorial or equivalent methods.
- Scanning in **300dpi greyscale** is on most scanners *exactly as quick* as scanning in 300dpi black/white or in any lower resolution! You will not save time if you scan in 300dpi black/white or in 200dpi instead of 300dpi greyscale, but you do lose *a lot* of quality.
- Scanning in **300dpi greyscale** produces large intermediate scanned files, which will be processed into very small DJVU files. Scanning in 600dpi black/white produces smaller intermediate scanned files, but the process of scanning at 600dpi is much slower for most scanners. Also, it's easier to process 300dpi greyscale scans because they have less "digital dirt" than 600dpi black/white scans.
- It is nearly impossible to improve the quality of a poorly scanned and/or incorrectly processed image of a book. For example, some e-books are made by inexperienced people in 150dpi, or in color instead of black/white, or the resolution was decreased after scanning in an attempt to reduce the file size. These e-book files are huge in size. The visual and print quality of such e-books is bad and **cannot** be improved! It is important (and not difficult) to make the scanned image correctly and ensure great quality of the resulting e-books. Read on!

A high-quality scanned e-book is small in size, has great visual appearance on the screen and also when printed, and has searchable text. There are many ways to achieve high quality of scanned e-books; all methods involve the resolution of **600dpi**. (Higher resolution almost never brings a significantly better quality.) Output files are in the **DJVU**[3] format and take typically about 5KB/page to 10KB/page. If your file is significantly larger, while the book contains only black/white text and is printed reasonably clearly, something was done incorrectly when producing the file.

You may of course experiment on your own with other programs. For example, some people use Photoshop with special plugins, Book Restorer, Corel PhotoPaint, RasterID, even Matlab and IDL for image processing. This tutorial presents a particular method that practically guarantees good results. If you are a beginner, please make a few books by closely following the instructions

[2] This kind of processing when the resolution of an image is increased is called *upsampling*.

[3] If you don't know what DJVU is, please use Google or Wikipedia to read about it. The DJVU format was specially developed for high-compression storage of scanned images. Most e-books today are in the PDF format, but the PDF format was intended for documents created in a word processor, i.e. for *vector* documents rather than *scanned* documents. Scanned e-books in PDF format occupy more space and/or display slower than in the DJVU format.

in this tutorial. You will then see that you can achieve quite a high a level of quality without excessive effort and without learning too many technicalities. If you develop your own methods, for example by using different options or different programs, you will be able to decide which method is best because you can then compare the quality of the results with the "reference" quality obtained by the methods in this tutorial.

2 Scanning a book

You pick up a thick volume. Maybe you think that only a maniac could scan it, page after page. Yes, you are right! But you can become that kind of maniac and scan books of any size without much discomfort if you organize your work well.

For the impatient reader:

- use any **flatbed scanner**, even a cheap one, and a program such as IrfanView to control scanning
- **do not** use a digital camera for scanning books!
- **do not** use FineReader for scanning books!

Why not use FineReader for scanning? The "FineReader" is a good program *for making OCR* but is not optimal for scanning and for processing the scans with the goal of making a scanned e-book. FineReader attempts to give you a kind of all-in-one solution for scanning and processing e-books; please resist the temptation to use just one program for everything. You will not get good results with FineReader; in any case, nowhere as good as when you follow this tutorial. FineReader has the following technical drawbacks: 1) It sometimes uses JPEG for image compression. This is not appropriate for black/white text! 2) It stores images internally as black/white 300dpi TIFFs and auto-rotates them. Black/white 300dpi is adequate for OCR but not optimal for digital scanned e-books. The auto-rotate algorithm is faulty and produces defects in the image ("broken" lines). The auto-rotation is hard-coded into FineReader 7.x, 8.x and cannot be disabled.[4] 3) If you scan in 300dpi greyscale, which is the procedure recommended here, FineReader will perform all operations at 300dpi, rather than resample to 600dpi. ScanKromsator and ScanTailor will first resample to 600dpi and then perform processing. The results of FineReader processing are always going to be inferior for these reasons.

Why not use a digital camera for scanning books? You will **never** get good results even with expensive 10 Megapixel or whatever cameras. Never even closely as good as with a flatbed scanner, even a cheap one. Look at figure 1 below and guess which of the two images of the same page is made by a digital camera.

[4] Only in FineReader version 9 there was added an option to disable this auto-rotation. However, other features of FineReader remain. Also not ethat FineReader version 9 cannot be used to produce OCR layer in DJVU files. I recommend using FineReader version 8.

Proof. First we will show how transition functions can be constructed for the given map (1.1.4) and vice versa. For the given point $x_0 \in M$ we choose n linearly independent columns of the matrix $P(x)$, and let $e(x)$ be $N \times n$ matrix formed by these columns. In the neighborhood U of the point x_0 we have $\operatorname{rank} e(x) = n$. Similarly, let $g(x)$ be an $n \times N$ matrix formed by n rows linearly independent in U. Then the $n \times n$ matrix ge is nondegenerate, and we define $e' = (ge)^{-1}g$. It is clear that

$$e'e = 1, \quad ee' = P(x)$$

in U (here 1 denotes the identity $n \times n$ matrix). As M is compact, there exists a finite covering U_i and the matrices e_i, e'_i, defined in U_i, satisfying the above equalities

$$e'_i e_i = 1, \quad e_i e'_i = P(x) \tag{1.1.5}$$

Transition functions $f_{ij}(x)$ can be defined by

$$f_{ij}(x) = e'_i(x)e_j(x) \quad x \in U_i \cap U_j.$$

Conversely, given transition functions, we can define the projector $P(x)$ as a block matrix, with $n \times n$-blocks $p_{ij}(x)$ as follows. Let $\{\rho_i(x)\}$ be a quadratic *partition of unity*, that is a collection of nonnegative functions, such that $\rho_i(x) \equiv 0$ outside U_i and

$$\sum_i \rho_i^2(x) \equiv 1. \tag{1.1.6}$$

Set

$$p_{ij} = \rho_i f_{ij} \rho_j, \quad i,j = 1,2,\dots,N. \tag{1.1.7}$$

Then, using (1.1.1), (1.1.2), we get

$$\sum_j p_{ij}\, p_{jk} = \rho_i \sum_j (f_{ij}\rho_j^2 f_{jk})\rho_k = \rho_i f_{ik} \sum_j \rho_j^2 \rho_k = p_{ik},$$

which means that the blocks p_{ij} give a projector matrix. The *frame* and the *coframe* matrices e_i and e'_i can be given as

$$e_i = \begin{pmatrix} \rho_{1i} f_{i1} \\ \rho_{2i} f_{i2} \\ \vdots \\ \rho_{Ni} f_{iN} \end{pmatrix}; \quad e'_i = (\rho_{i1} f_{1i}, \cdots, \rho_{iN} f_{Ni}).$$

It is easy to verify the identities (1.1.5) for these matrices. This completes the proof of the first part of the theorem.

For two isomorphic bundles E, E' with transition functions f_{ij}, f'_{ij} define the projectors P, P' by (1.1.7). Using lemma 1.1.3, we may wright

$$P' = A\,R, \quad P = R\,A, \tag{1.1.8}$$

14 1 Elements of Differential Geometry

Proof. First we will show how transition functions can be constructed for the given map (1.1.4) and vice versa. For the given point $x_0 \in M$ we choose n linearly independent columns of the matrix $P(x)$, and let $e(x)$ be $N \times n$ matrix formed by these columns. In the neighborhood U of the point x_0 we have $\operatorname{rank} e(x) = n$. Similarly, let $g(x)$ be an $n \times N$ matrix formed by n rows linearly independent in U. Then the $n \times n$ matrix ge is nondegenerate, and we define $e' = (ge)^{-1}g$. It is clear that

$$e'e = 1, \quad ee' = P(x)$$

in U (here 1 denotes the identity $n \times n$ matrix). As M is compact, there exists a finite covering U_i and the matrices e_i, e'_i, defined in U_i, satisfying the above equalities

$$e'_i e_i = 1, \quad e_i e'_i = P(x) \tag{1.1.5}$$

Transition functions $f_{ij}(x)$ can be defined by

$$f_{ij}(x) = e'_i(x)e_j(x) \quad x \in U_i \cap U_j.$$

Conversely, given transition functions, we can define the projector $P(x)$ as a block matrix, with $n \times n$-blocks $p_{ij}(x)$ as follows. Let $\{\rho_i(x)\}$ be a quadratic *partition of unity*, that is a collection of nonnegative functions, such that $\rho_i(x) \equiv 0$ outside U_i and

$$\sum_i \rho_i^2(x) \equiv 1. \tag{1.1.6}$$

Set

$$p_{ij} = \rho_i f_{ij} \rho_j, \quad i,j = 1,2,\dots,N. \tag{1.1.7}$$

Then, using (1.1.1), (1.1.2), we get

$$\sum_j p_{ij}\, p_{jk} = \rho_i \sum_j (f_{ij}\rho_j^2 f_{jk})\rho_k = \rho_i f_{ik} \sum_j \rho_j^2 \rho_k = p_{ik},$$

which means that the blocks p_{ij} give a projector matrix. The *frame* and the *coframe* matrices e_i and e'_i can be given as

$$e_i = \begin{pmatrix} \rho_{1i} f_{i1} \\ \rho_{2i} f_{i2} \\ \vdots \\ \rho_{Ni} f_{iN} \end{pmatrix}; \quad e'_i = (\rho_{i1} f_{1i}, \cdots, \rho_{iN} f_{Ni}).$$

It is easy to verify the identities (1.1.5) for these matrices. This completes the proof of the first part of the theorem.

For two isomorphic bundles E, E' with transition functions f_{ij}, f'_{ij} define the projectors P, P' by (1.1.7). Using lemma 1.1.3, we may wright

$$P' = A\,R, \quad P = R\,A, \tag{1.1.8}$$

Figure 1: Two images of the same page, one made by a digital camera, another by a cheap flatbed scanner. The image made by a flatbed scanner was scanned at 300dpi greyscale and upsampled to 600dpi black/white. You can guess which image was made by the digital camera! (Yes, the crappy one.) We recommend that you always use a flatbed scanner and scan at 300dpi greyscale or higher resolution.

For scanning, you need basically any program that can work with your scanner. Under Windows, the TWAIN scanner driver is popular.[5] Under Linux, many scanners are supported by the VueScan program, but you can use any other program as long as your scanner is supported.

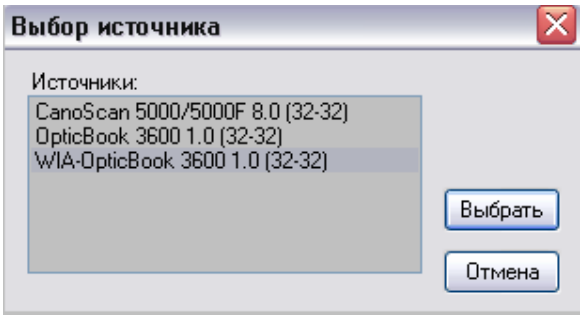
You can scan using any program like IrfanView, XnView, ACDSee, PhotoShop. (Note that IrfanView is small an free.) It is important that your scanning program does not try to do anything with the images; in particular, no deskew, no “optimizing”, no resizing, nothing at all. You should be able to tell the program just to save the scans for each page to the hard disk in the TIF format.

It is convenient if your scanning program can save scanned images for every page one after another, numbering the files like p0001.tif, p0002.tif, etc. For example, VueScan and IrfanView can do this.

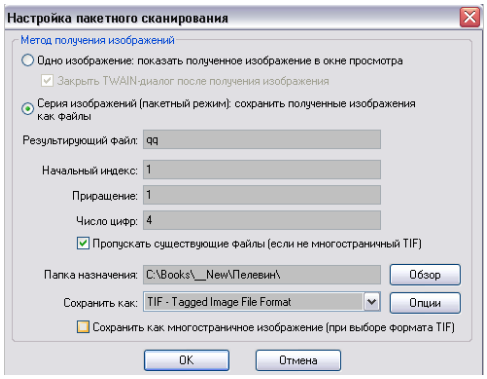
2.1 Setting up IrfanView for scanning

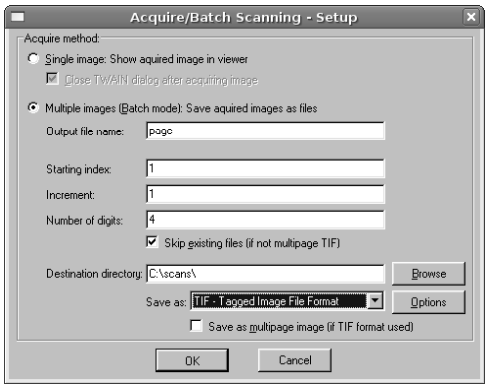
As an example, we describe how to scan using IrfanView. (This program can be downloaded for free.) Scanning in other programs is quite similar.

Start IrfanView. In the File menu, press "Choose TWAIN Source". Choose the scanner that you need to use.



Then in the same menu choose "Acquire/Batch scan".

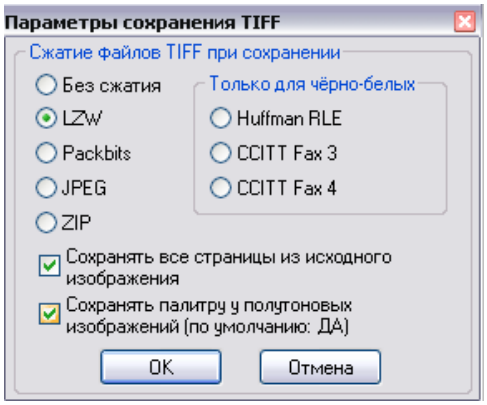


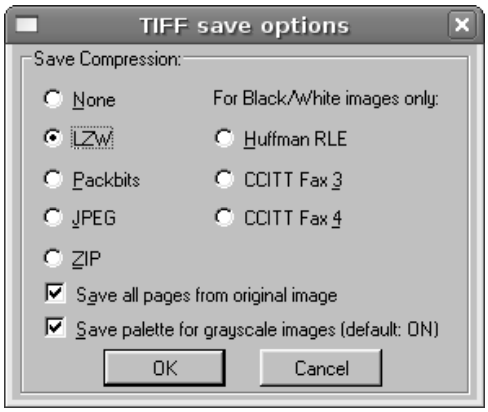

Acquire/Batch Scanning - Setup
Acquire method:
Single image: Show aquired image in viewer
Close TWAIN dialog after acquiring image
Multiple images (Batch mode): Save aquired images as files
Output file name: page
Starting index: 1
Increment: 1
Number of digits: 4
Skip existing files (if not multipage TIF)
Destination directory: C:\scans\
Browse
Save as: TIF - Tagged Image File Format
Options
Save as multipage image (if TIF format used)
OK
Cancel

[5] Most scanners are supported by TWAIN drivers; for other scanners you may need special drivers.

Here you can choose how to number the scanned files, where to store them, and in which format to save them. As shown, the files will be named page0001.tif, page0002.tif, etc. You should select **TIFF** as the image format. (**Do not** use JPEG as the output format!)

Click on Options to the right of "Save as" field. This will set the options for the TIFF format.





You should select "LZW" compression; this will cut the TIFF file size in two, compared with no compression ("None").[6] If you later find that you have compatibility problems with these TIFF files (i.e. you later use a program that cannot open them) then you need to change the compression method.

Important: **Do not** use the JPEG compression method for black/white text! JPEG compression introduces *digital artifacts*, that is funny-looking shades around each letter (see figure 2). It is pointless to use JPEG for black/white images.[7]

Now press OK and go to the TWAIN driver window for your scanner.

In the TWAIN window (or other configuration window if you are not using TWAIN drivers), set the resolution to **300dpi** and the color mode to **greyscale**. (In some programs, this is called "8 bit greyscale".) These are the most important settings. Some scanning programs do not allow you to set explicitly the resolution or the color mode; instead they say something like "Black/white photo" or "web-optimized quality". Avoid these programs, instead use some program that allows you to set specifically 300dpi and 8-bit greyscale. If you are not sure that your settings are right, you should try scanning one page, save the file to disk as TIF, and check the properties of the file in a graphics editor, to make sure that you actually got 300dpi and 8-bit greyscale.

[6] Note that a typical page scanned in greyscale will occupy between 2 and 4 megabytes on the hard disk with LZW compression.

[7] The JPEG format actually cannot handle black/white images; when one converts black/white images to JPEG, the software must convert those images into greyscale images. The JPEG compression then introduces a certain quality loss, as shown in the figure. The quality loss in JPEG compression is acceptable for photographs but may degrade black/white text quite significantly, unless a high quality JPEG mode is selected. (The quality of JPEG compression is usually selectable as a number from 1% to 100%. No visible artifacts would appear at 90% quality or higher. But some programs, especially for making PDF files or for "optimizing" images, may not allow you to set the JPEG quality manually.)

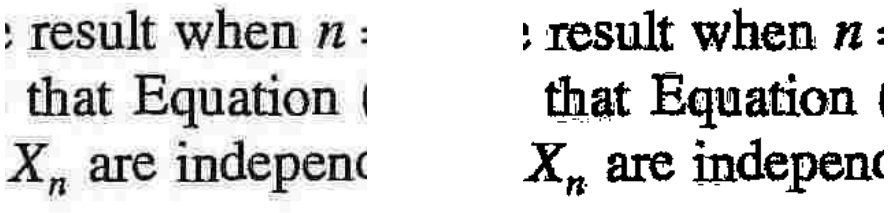


Figure 2: Digital artifacts appearing due to JPEG compression of black/white text. (In this example, the quality setting for the JPEG encoding was very low, so these artifacts are apparent to the eye.) At left: greyscale image with unnatural wavy-looking shadows around the letters. These "digital shadows" are typical for JPEG compression of black/white images. At right: the same image converted back to black/white. The digital artifacts produce "digital noise" i.e. speckles around the letters and a distortion of the shape of the letters.

2.2 Setting up VueScan for scanning

VueScan runs under Linux, Windows, and Macintosh, is not a free program, but all upgrades are free once you buy it. An advantage of VueScan under Linux is that it supports many types of scanners that are otherwise not supported by standard Linux software.

In VueScan there are many tabs with options. The first tab (figure 3 left) is the "Input" tab that controls the scanning mode. Note that VueScan may not show you all these options unless you enable the "Expert" mode (or "show all options"). You can make the settings as shown; for instance, you explicitly set the resolution to 300dpi and the color mode to 8 bit greyscale. It is important to check the box "Lock image color" so that each page is scanned in the same color balance. If you want, you can make automatic scanning with a small delay; then you will have to jump to the scanner every time to change the page. I prefer not to do this.

The second tab (figure 3 left) is the "Output" tab. There you can set the directory where the scans will be kept, the format of file names (in this example, it will be p001.tif, p002.tif, etc.), and the TIFF compression ("On").

2.3 Handwork while scanning

By now you have set up your scanning program. The actual work while scanning is not complicated:

- First you need to try scanning some place in the book and check that everything works well. Take a book, open somewhere where the pages are full of text, put the book (both pages down) on the scanner glass.
- If necessary, press with your hand so that the crease of the book is as close to the glass as possible. (You can also use a weight, e.g. put another heavy book on top, but it's slower than pressing by hand.)

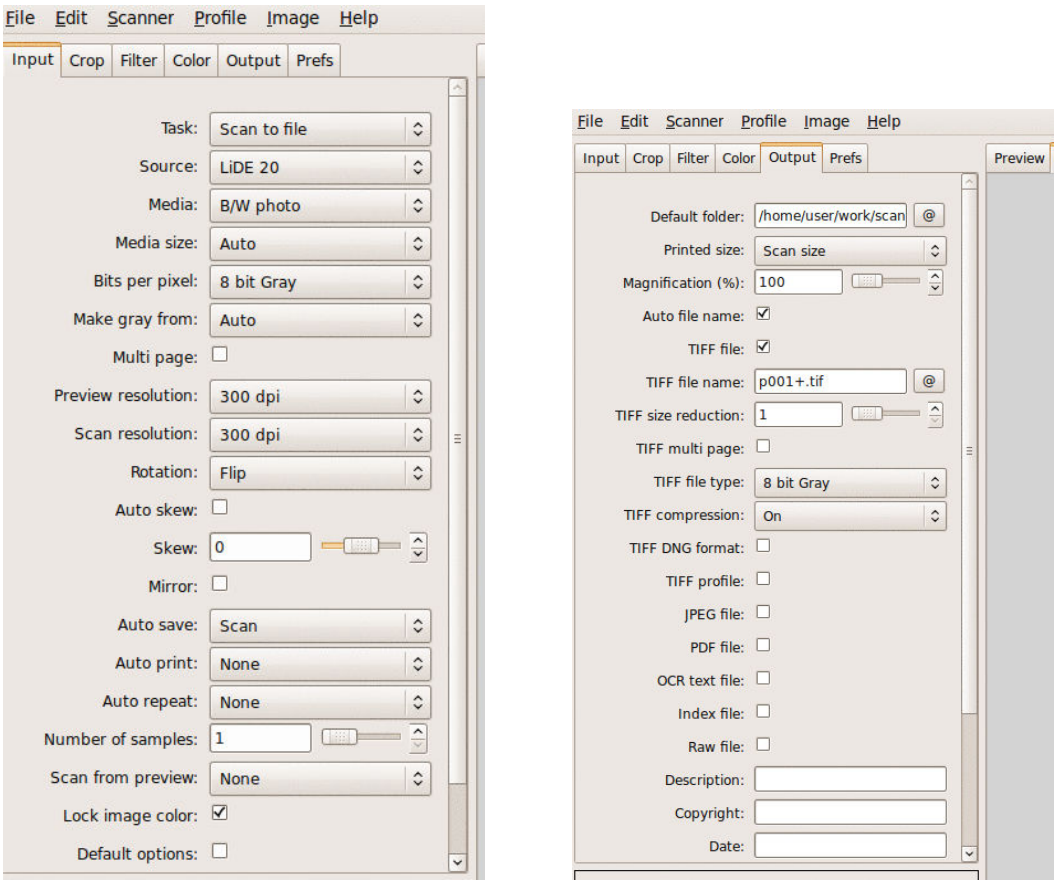


Figure 3: Options for scanning when using VueScan.

- Do a "preview scan." Then you can see what has been scanned in the preview window. If needed, you can turn the page 90 degrees so that the text is straight up. You can also adjust contrast, brightness, gamma correction if necessary. Your goal is that the text must be clearly visible, not too dark, and not too light.

- Select the scanning region by using the mouse. You should select the scanning region such that some white space is left around the text, but no book crease or off-page regions are scanned. Your purpose is that the scanning rectangle should fit around the text with some margin, so that you will not lose any text even if you put the book a little askew on the glass. And yet you do not want to scan any useless regions outside of the page.

- Press the "Scan" button with the mouse and wait until the scanner finishes scanning the page. This will get the scan of one page (or two pages at once, if you can fit the book onto the scanner). The scanned file will be saved to the disk.

- Now that the scanning program is set up, you can scan all the pages with the same settings. While the scanner lamp is moving back, turn the next page and put the book back to the same place on the scanner. Then press the mouse button to scan again. (The mouse can be left pointing at the "Scan" button, so you don't need to look. Alternatively, some scanners have buttons on them that make the next scan.)

This technique allows you to scan the entire book, one page after another, without looking at the computer screen or at the keyboard. You can watch TV or whatever while you are scanning. Depending on the scanner speed, you can get between 100 and 200 scanned pages per hour. Some scanners are particularly fast (e.g. Plustek OpticBook).

It is not necessary to set the book onto the scaner absolutely straight (edge of the book parallel to the edge of the scanner). You should try to put it reasonably straight, but it is unavoidable that pages will not all be scanned completely straight; many pages will be *slightly skewed*. This small skew is okay and will be corrected later (after scanning) by software. Correcting this skew is called *deskewing*. Deskewing is very fast and efficient.

What you want to avoid when scanning:

- Avoid very large skew angles, i.e. do not place a book at a large angle on the glass. This kind of scan can still be deskewed, but the shapes will probably not be as smooth as otherwise.

- Avoid incomplete page scans, i.e. when some of the text is outside of the scanning region. This means that some text will be lost (not scanned at all). If you discover such a page, scan that page again with a correct scanning region. In a science book, no part of the text is unimportant!

However, avoid scanning the library stamps or other marks on the pages. If your book has stamps or other markings on some pages, just cover them with a piece of paper while scanning, or remove them with digital image editor after scanning. Nobody wants to see some ugly stamps or marks in the e-book!

- Avoid scanning any off-page regions (this will be when your scanning rectangle is way too large). This will produce a black shadow which, in many cases, you will have to remove by hand while processing your scans! (This is so because computers are not very good at guessing what is a part of the book and what is "dirt" on a scan.)
- Also, avoid producing a fuzzy image because some place on the page was not close to the scanner glass.

The region of the text around the book crease is often difficult to scan. You can try scanning one page at a time (rather than two pages) or pressing slightly harder onto the book binding. It is important that the text is directly next to the scanner glass. Even 1 mm distance between the glass and the paper will make a very fuzzy scanned image in almost all scanners! Fuzzy scanned images are *not acceptable*. It is very difficult to prepare a good quality final e-book from fuzzy scans.

Should you scan one page at a time, or two pages at a time? It is faster to scan a book two pages per scan rather than one page at a time. Double-page scans can be cut quite efficiently and automatically (if they are scanned cleanly) by software. But not all books can be scanned that way; many books are too large (you won't fit two pages onto the glass unless you have an A3 scanner, which is usually expensive). Many books don't open sufficiently to be scanned two pages per scan with good quality (some text near the crease is lost or becomes too fuzzy, which is *not acceptable*!). You need to try two pages at a time, try one page at a time, and then decide how to proceed. Regardless of how you scan, the processing software will be able to prepare an e-book with single page images, as long as everything is scanned correctly and the images are clear.

The result, after scanning the entire book, is a directory full of TIFF files. These files are the *raw material* that you will start processing after you finish scanning. Note that you need sufficient disk space to store all those scans (at least 4MB per scanned image!). After you finish scanning, use a slideshow mode of some picture viewer to quickly preview the scanned images to make sure that you didn't miss any pages and that every page is adequately scanned. It will be too late when you discover, at the final processing stage, that some pages are only half-scanned or missing, especially when the book has already left your hands!

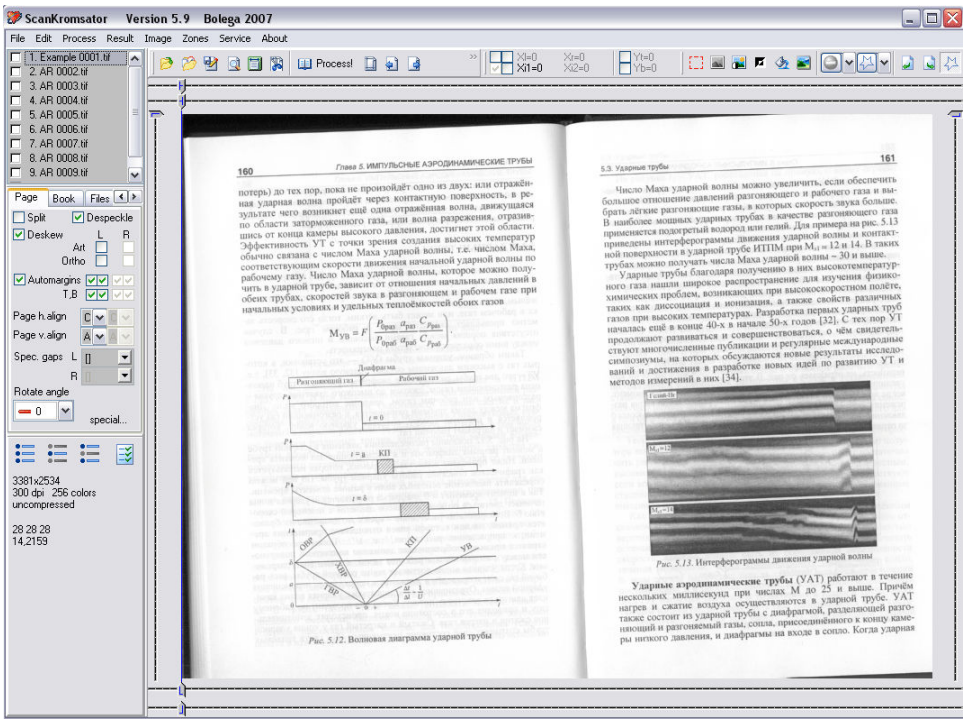
Note: When you scan the book, scan *all* pages; please *do not omit* any pages, including title pages, front matter, including any information about the publisher, the table of contents, the index, the bibliography, empty pages in the middle of the book, page numbers, errata sheets, or anything else!!! You will

not save much time if you decide to skip 20 pages or so while scanning. However, a science book is almost unusable without bibliography and index and without exact information about its publication. [8]

3 Processing scans with ScanKromsator

Now we discuss the processing software. The first program is the wonderful ScanKromsator written by Bolega.[9] ScanKromsator is a very powerful tool for processing scanned material. ScanKromsator has a very large number of useful functions, but some of them are not intuitive or difficult to understand if you just look at the user interface.[10] In this tutorial you will be walked through a particular simplified workflow with ScanKromsator, assuming that you scanned a book at 300dpi greyscale.

Start ScanKromsator and load the raw TIFF files into it (menu File). The list of files will appear on the top left column. The **toolbar** with several **tabs** ("Book", etc.) will appear below the list of files.



In the example shown, a book was scanned with two pages per scan, and apparently there was some skewing. Our task now is to split, to deskew, and

[8] Also, do not think that you will make your life easier from the legal point of view if you don't scan the publication information!

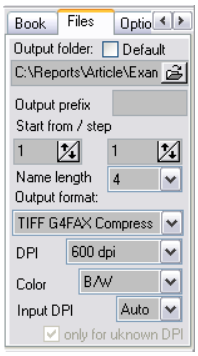
[9] Please do not write email to Bolega asking for help, for documentation, for source code of ScanKromsator, or for adding extra features! Instead, just learn to use it and make some good quality e-books!

[10] We will talk only about the bare minimum of ScanKromsator functions here. Unfortunately the ScanKromsator program does not yet have a comprehensive user's manual describing all the functions.

to cut the page images so that every page has the same size and margins. If your scan is single-page, you will not need to split, but you will still need to deskew and cut. This operation is called “kromsating” in the program.[11]

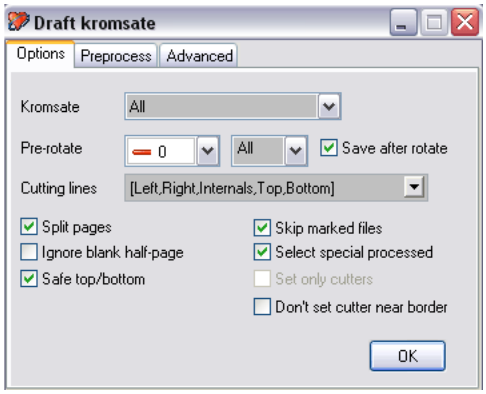
3.1 Draft run

The first step is a *draft* processing run, i.e. preparation for the final processing of the raw files.



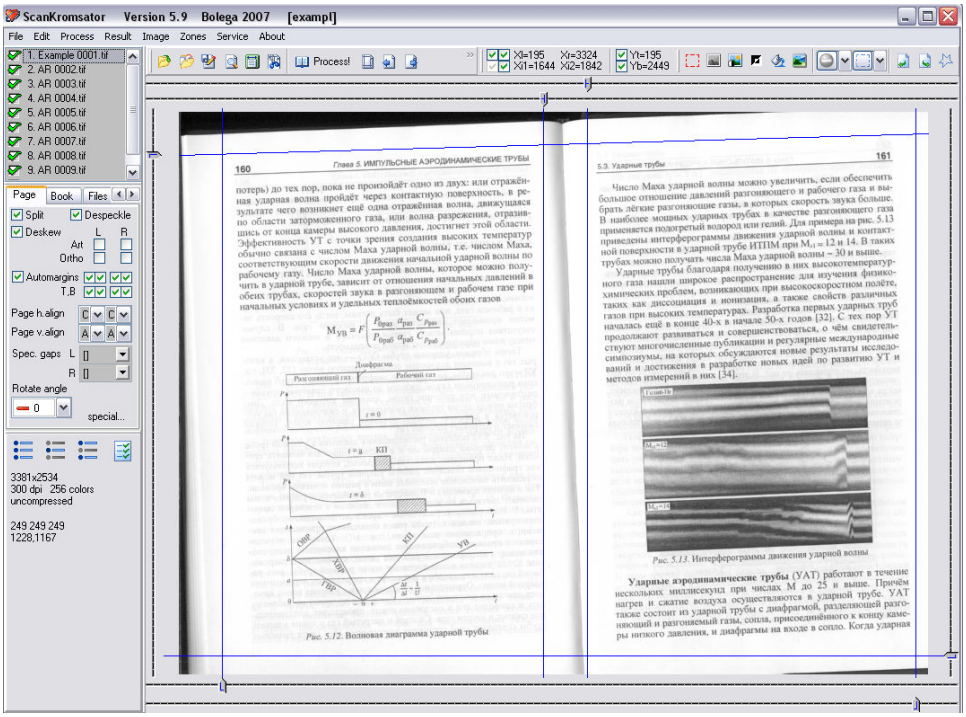
Click the tab “Files” **in the toolbar**. You get a dialog where you can set the output resolution (very important!) to 600dpi, the folder for storing the output files (the output folder is by default the subdirectory `out` in the current directory), and the way of numbering the output files (prefix, number of digits, starting number, step). Note the format for compressing the output files: it’s TIFF G4 encoding, which is optimal for black/white TIFF images. This will be the output format after processing.

To start the draft processing run, click the button “Draft kromsate” bearing the pictogram of scissors, which is located to the left of the “Process” button in the toolbar. When you press the “Draft kromsate” button, and you get the dialog shown at right. In this dialog you need to set tick marks on “Split pages” and “Safe top/bottom.” The field “Kromsate”=All means that the options are applied to all the pages. If some pages do not need to be split, you can select “Kromsate”=Current and unset “Split pages” for these pages.



Press OK and wait 10-15 minutes until the “Draft kromsate” operation is finished. You will get the following screen.

[11] The pseudoword “kromsate” is a mangled Russian word meaning “to cut in pieces.” Within the ScanKromsator, the meaning of “kromsate” is the operation of splitting a two-page scanned image into individual page images, and also the operation of cutting page images so that the margins become even and equal on all pages.



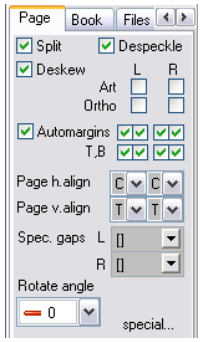
Note that there are now green tick marks in the page list (top left column), meaning that these pages have been “draft kromsated” successfully. For each page you will see the blue lines across the page. These lines are the *cutters* that determine how the page image will be cut and split. Note that the program attempts to determine automatically where to cut the margins and where to split a two-page image into single pages. In some cases the program may make a mistake and cut too much or too little; in that case you will later be able to adjust the position of the cutters by hand.

3.2 Set options

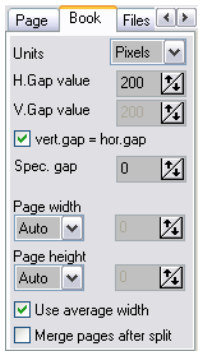
The next important step is to go through the processing options and prepare for the main (not “draft”) run of ScanKromsator. The processing options are set in the many different tabs in the toolbar (left middle column).

Please note: Each option can be set either to apply to all pages at once, or only to the currently shown page. To apply an option to all pages, hold the **Ctrl** key while clicking the option box with the mouse. In this way, you can set some common options quickly for the entire task and then go to some problematic page and select other options just for that page.

First click the “Page” tab. Here you can set processing options for cutting the pages. The option “Split” means to split the two-page image into single pages. “Deskew” will deskew each single page image separately. “Despeckle” removes small dots. Sometimes “Deskew” makes pages significantly skewed; this is usually due to some complicated illustrations. In that case, check “Art” for these pages. You can set “Ortho” if the page needs to be rotated by 90 degrees. You can set these options separately for left and right (L and R) pages.



Now click on the “Book” tab. Here you set options related to the size and layout of the pages in the final book. “H.Gap” is the size of the margins. The value of 200 is good for 600dpi (meaning 1/3 inch). Page width and height can be set to Auto. You can also center the pages differently (align to center/align to top/align to bottom).

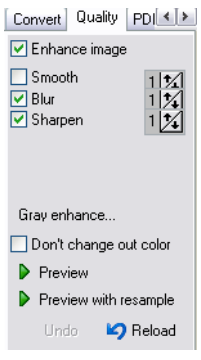


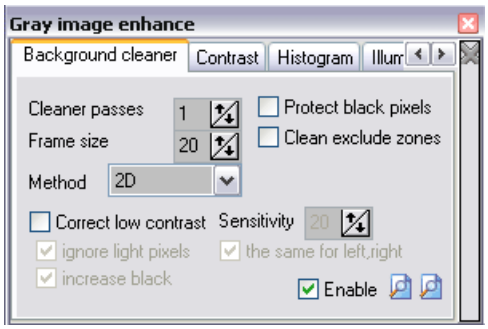
We already visited the “Files” tab at the “draft” stage. It is **very important** to have 600dpi as the output resolution in the “Files” tab!

Now click on the “Options” tab. Set “Deskew method” = Auto (shear), Resample filter = Lanczos3. The setting “Despeckle”=Fine+Normal or Safe switches on an “intelligent” despeckle method that avoids removing the dots over i or j, for example. “Text sensitivity” controls the logic of the autocutting. Low sensitivity might cut off the page numbers if they are too far away from the text. You may need to adjust the sensitivity settings a little bit; but in most cases they do not need to be adjusted.

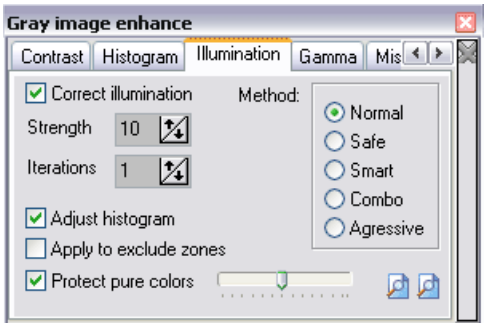
You can skip the “Options 2” tab for now. Click on the “Convert” tab. Here you set the threshold for converting greyscale images to black/white. Do not forget to hold the Ctrl key (to set this for all pages) as you select “Threshold”=MiddleDark. Experiment with other settings if you don’t like the results.

Click the “Quality” tab; there you can further control the conversion to black/white. This is a very important function! Set Enhance image, Blur=1, and Sharpen=1. What is important is that the image will become smoother with this setting. The values of Blur and Sharpen could be 2 instead of 1, although the value 1 is usually good. A larger value will make the letters more black. You may need to experiment depending on the quality of printing in a particular book.
Another important option is “Gray enhance.” Click on it since you have greyscale scans (which is what you should have!).

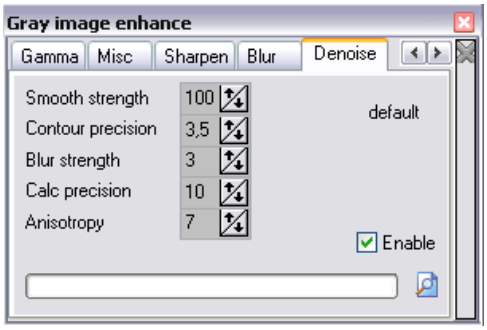




You will get a dialog with many options for greyscale images. Go to the "Background cleaner" tab and check "Enable".



Skip several tabs and click the "Illumination" tab; click "Correct illumination". This will normalize the illumination of the page, which is important since usually some parts of the page are darker than others. This is a very useful feature that removes black shadows that would otherwise appear in darker places on the page!



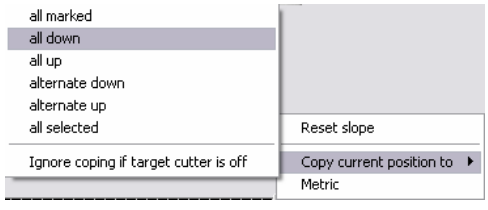
Skip several tabs and click the "Denoise" tab. Set the parameters as shown at right. These parameters clean up the image. This is the last set of options that we are going to bother with right now.

You can use the File→Options... menu to write the options to a file. This will save you all this work for the next time.

The last step before the main processing is a visual checking of the position of the cutters. You need to go through *every* page and check that the cutters are correctly positioned. Yes, this is a bit boring... but you can make it quick.

Put two fingers of the left hand onto the keys **q** and **w**; pressing these keys will go to the previous/next page. With the right hand, you hold the mouse and adjust the position of the cutters wherever needed. Sometimes there is a skewed shadow, or it is necessary for some reason to set the cutter line *at an angle* rather than vertically or horizontally. Hold the **Shift** key and drag the cutter by its end to achieve this.

You can copy the cutter position from one page to another. Right-click on the cutter, and you will see the menu as shown. For instance, if the current cutter position needs to be applied to all subsequent pages, click "Copy current position to"→"all down."



If some page contains a photograph or a color figure, you need to protect it from converting to black/white. This can be done when checking the position of the cutters. Basically, you can select some arbitrary part of the page and mark it as a *picture zone*. See Section 3.4 for more details.

You can save the settings for this task by using the File/Save Task command in the menu. This command is useful if you want to stop the task and to continue it later.

3.3 Main run

Now that everything is ready, you can begin the main run of ScanKromsator. Press the large button that says "Process" and bears the icon of a book, in the main toolbar at top:

The program will ask you to confirm that you really are sure you want to change the resolution of the images. Confirm! The process will then start.

Now you need to wait a while. The upsampling operation can be quite slow; in recent versions of ScanKromsator (5.8 and up) this operation was made faster. You may expect to process 5 pages per minute or so. When everything is finished, you should view the output files in the output folder. You should check that all pages are cut and deskewed correctly. If some pages are not processed correctly, you can repeat processing of just those pages with some other options.

The main processing run may take some hours on a slow computer. It is not necessary to process the entire book in one run. One can process only some portion of the pages; then one needs to set Book→Page width→Fixed to the size determined in the previous portion of the pages (so that all pages have equal size at the end of processing). It is usually sufficient to take 10 to 15 pages for determining page size.

If you like, you can use the powerful cleaning features of ScanKromsator to remove the "digital dirt" from some pages. Typically, the "digital dirt" is any extraneous spots on the paper, pencil or pen marks, and library stamps. Of course, you can also use any graphics editor to clean the images by hand. Hopefully, there will not be many pages to clean.

3.4 Processing color figures and photos

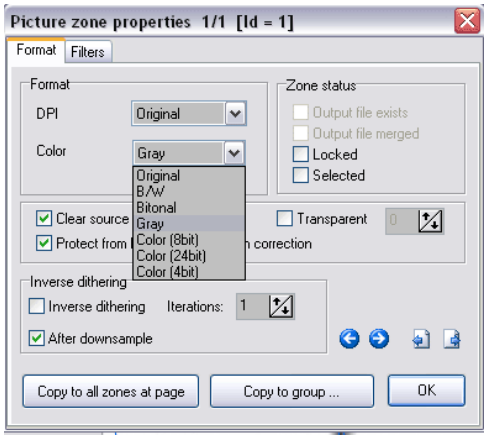
We discuss color figures separately because they are not frequently needed. However, their place in the workflow is at the point where you check and adjust the position of the cutters.

The latest version of Kromsator (5.9) includes a feature for color figure processing, the so-called *picture zones*. One some pages there may be a *picture*,

i.e. a non-black-white illustration such as a photograph or a colorful diagram. You need to protect these illustrations from converting into black/white. To mark a picture zone, select a rectangle containing the illustration and click on the button “Mark as picture zone” bearing the icon of a blue frame in this toolbar:

There is also a possibility to have polygon-shaped picture zones. This is useful, for example, if the page was scanned with a large skewing. Use the star-shaped tool button to mark such zones:

To set the options for a picture zone, double-click on the selected region. You will see the dialog “Picture zone properties.”



You need to set the color of the illustration. For example, if the page contains a greyscale photograph (rather than a color photograph or color diagram), set Color=Gray.

We cannot discuss other zone options here; as you see, there are many options intended for advanced users. But note that after “kromsating” the picture zones will be saved to *separate files*. So after the main processing run you will have to merge them with the page files. This is done by using the menu command: Zones→Picture zone→Merge zones. The resulting page files will be TIFF files in which the text is black/white but the picture zones have color.

4 Processing scans with ScanTailor

ScanTailor is a relatively new program that is being actively developed; I describe version 0.9.8 at this time. It can be downloaded for free and runs under Windows and Linux.

The functionality of ScanTailor is sufficient for processing books that have black/white text and some greyscale illustrations, as well as occasional color pages. ScanTailor can deskew and clean up your scanned pages, split double-page scans into single pages, and convert from 300dpi greyscale into 600dpi black/white, while keeping greyscale illustrations.

ScanTailor has online documentation at its website; you can read about many features of ScanTailor there. Therefore, here I will only show how to do the most common processing steps.

4.1 Importing scan into ScanTailor

ScanTailor takes as input a number of TIFF files, and produces as output a new set of TIFF files. When you run ScanTailor, it first asks you to start a new "project" or to open a previous "project". A "project" means a bunch of TIFF files that are going to be processed together. So you say "new project" at this point (figure 4).

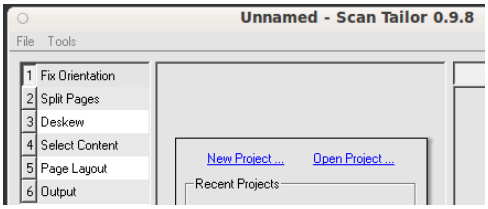


Figure 4: ScanTailor asks to create a new project or to open a previously existing project.

Then you will see a dialog box asking you to select the input files (figure 5). Press "Browse" on "Input directory" and select the directory where you have your scanned TIFF files. (The output directory will be automatically selected as the "out" subdirectory. For example, if your scans are in C:\myscans\ then the output TIFF files will be in C:\myscans\out.) You can now use the arrow buttons ("<<" and ">>") to exclude some of the TIFF files from processing. You probably want "Select all" at this point (i.e. use all the TIFFs in that directory). Then press "OK"; the TIFF files will be inspected and a "project" will be created. If some of the TIFFs do not have the correct resolution stamped inside them, you can correct it ("fix DPIs"), but normally this is not necessary.

After this, you see the main window of ScanTailor that looks like the following:

The selected page is shown in the central window; thumbnails of all pages are shown in the column at right; and the processing sequence (which I will explain shortly) is shown on the left.

ScanTailor's "projects" can be saved to files with the extension ".scantailor"; these files are in the XML format and have the full information needed to process the input TIFF files from the project and to produce the output files. So it is advisable to save the project, also while working in ScanTailor. So you go to the File menu and choose "Save" and specify the location and the name of the project file; for example "myscan".

It is also advisable to make the ScanTailor window maximized to full screen; but I will keep this window small in my examples, just to make screenshots smaller in this PDF file.

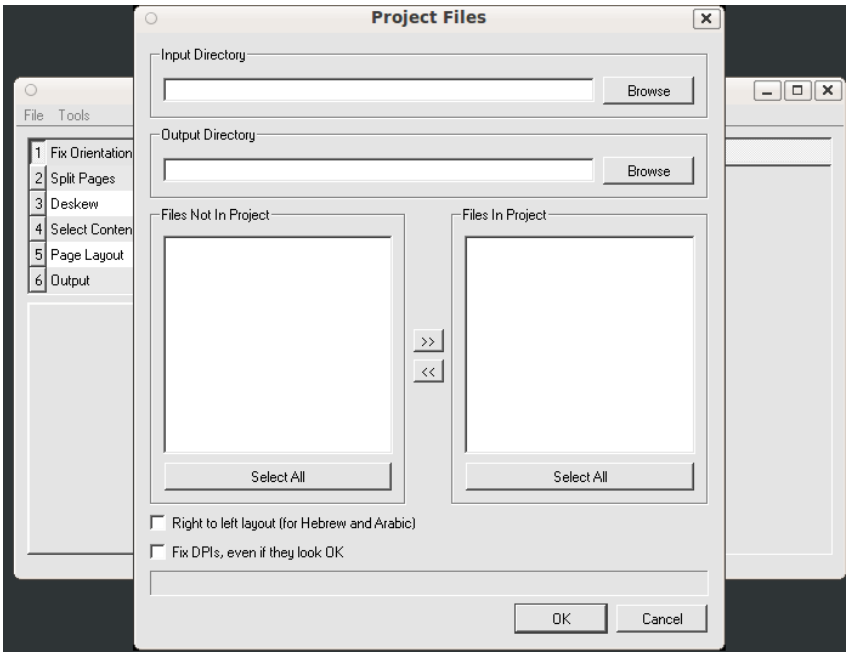


Figure 5: ScanTailor asks you to select the input files.

4.2 Draft run

Now that your scans are loaded into ScanTailor, you can start processing. The optimal way of processing is to let ScanTailor run automatically for all pages and then correct the errors that may have been made. Even when the scanned material is very simple and no user interaction is really needed, it is necessary to have a "draft run" and a "final run" because the final output cannot be produced until all final page sizes are known, and the page sizes are computed only after the "page layout" step is performed on all pages.

For the draft run, I suggest the following procedure that seems to be quickest:

- You already have the first page selected when you open the project for the first time. Press with the mouse on the "Page layout" step (or simply press "P" on the keyboard) and wait a little. The first page will be processed through all the steps 1-5 and then you will see the page layout dialog (figure 7). You will see that the first page has been really processed: deskewed, split, and a content rectangle was selected (everything outside the content rectangle has been cut away). Don't worry about any options at this point.
- If your scans contain double pages, automatic splitting is what you need. But if your scans contain *only* single pages and never need to be split, it is perhaps better to disable splitting. Click on "Split Pages", then click on "Change...", then select "Mode: Manual" and "Scope: All pages". This will effectively disable automatic splitting for all pages. Now click again on "5 Page Layout".
- Now press the "play" button to the right of "5 Page Layout".

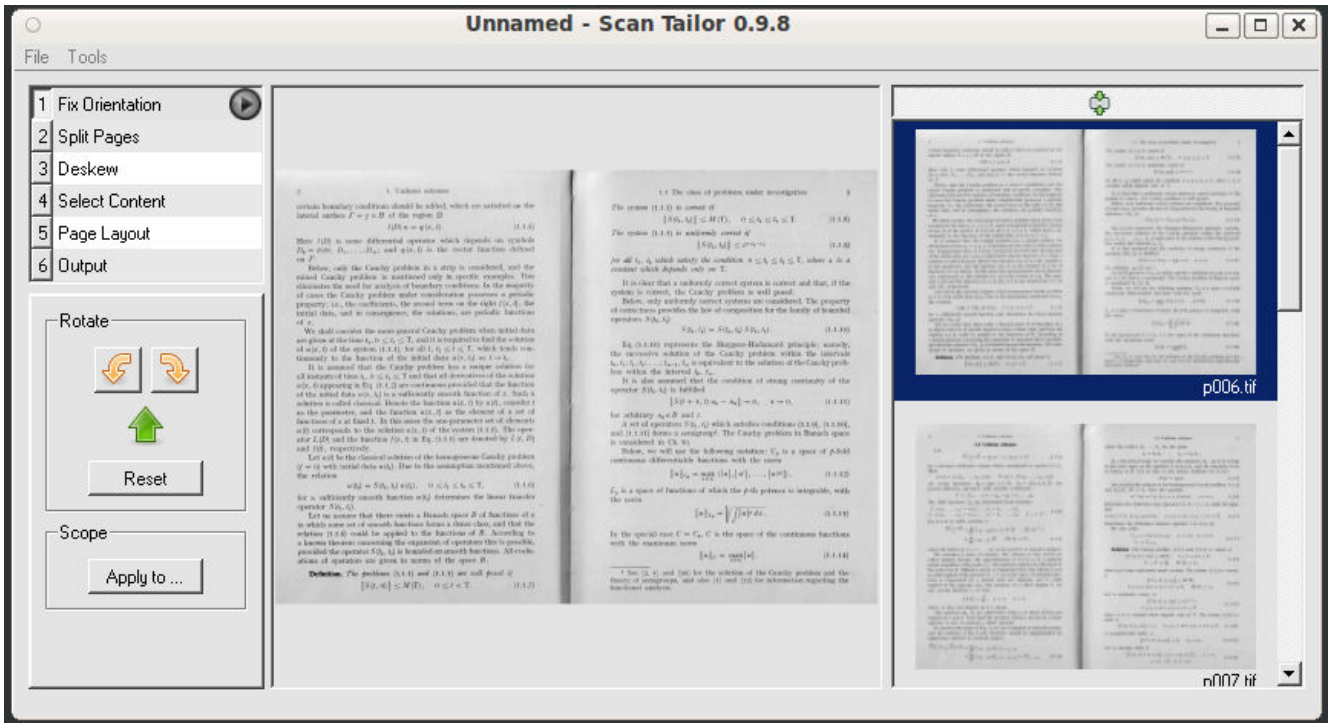


Figure 6: ScanTailor's main window with some scans loaded into a project "Unnamed".

This will start the automatic ("batch") processing of steps 1-5 for all pages with the default options. This process will take maybe 20 minutes or so (maybe about 5 seconds per page), but at least you don't have to do anything while the program is working. This is your "draft run." While it is running, let me try to explain what is actually happening now.

4.3 More about processing steps

The idea of ScanTailor is to divide the processing into steps as shown in the left of figure 6. Each step requires that all the previous steps are already performed on a given page. There is (in version 0.9.8) no way to omit some steps entirely from processing. You will have control over each step of the processing and can in principle adjust the settings for each page separately or apply special settings to a group of pages.

The first step is "fix orientation". Here you can rotate pages by 90 degrees or by 180 degrees, so that the text on the pages is more or less upright. This step is completely *manual*; the user needs to supply the rotation for each page or for all pages. (In order to apply some option for all pages, usually you need to press the "Apply" button and then select "Apply to all pages".) You will not need to control this step at all if you adjusted the page orientation correctly while scanning. By default, ScanTailor will not do anything at this step. However, you may go to a single particular page (choose it by clicking on the thumbnail in the right column) and change the orientation if needed.

The second step is "split pages." In the example shown in figure 6 there are

double-page scans that are already correctly oriented. Most likely, ScanTailor will *automatically* and correctly split them into single-page scans. In some rare cases the splitting is done incorrectly (e.g. too much text is cut off). In this case you can go back to the "split pages" step and correct this by hand. If your scans never need to be split, you can disable splitting (set it to "manual" for all pages).

The third step is "deskew", that is, a small rotation of each page to make the orientation completely upright. Note that deskewing is applied separately to every page, also to every split page. In most cases ScanTailor will correctly make the orientation of the text as horizontal as possible. In very rare cases you will have to adjust the deskewing by hand.

The fourth step is "select content". It selects the rectangle that seems to contain all the text on the page. *In quite a few pages* this rectangle will be too small or too big! (This is because it is difficult for the computer to understand automatically what the "actual text" is and what is some artifact of scanning, like a dark shadow at the edge of the page.) So it is at this step that you *certainly* will have to *look at every page* and check that the rectangle is selected correctly. More about this below.

The fifth step is "page layout". This step is fully manually controlled by the user; each page's "content rectangle" is aligned (if desired) with the content rectangles of all other pages, margins are added, and the resulting rectangle is prepared.

Since it is only at step 4 that problems are quite likely to appear while step 5 is completely manual, I propose to run all the steps 1-5 automatically as the "draft run". After the draft run, you will have to return to step 3 and flip manually through all pages to check that all is well. If needed, you will be able to return to *any* previous step for every page where that step produced an incorrect result. As experience shows, a non-negligible amount of work is needed only for step 4 at this point.

The last step is "output". At this step, which is usually quite slow but does not require any attention from you, ScanTailor will produce the resulting TIFF files in the output directory. After this step, you should flip through the final page images again, and check that everything is okay (especially if there were any color illustrations, see below). If there are no color illustrations, the output is usually fine without any further manual work.

It is important to understand that your original scanned TIFFs will never be changed; ScanTailor will only produce some new TIFFs in a different directory, and this will be done only at the last step (the "output"). These TIFFs will be the result of the ScanTailor processing.

4.4 Correct errors after the draft run

While the batch run is executing, you see the big "stop" button in the middle of the ScanTailor window. You can stop the automatic operation of ScanTailor at any time, by pressing on this "stop" button. Or you can "save" the project

file, also at any time *without stopping* the automatic run. This will save the information gathered up to that point. (What if the power is cut to your computer? Then you will be able to continue right from the point where you last saved the project file.)

When the draft run is completed, ScanTailor will stop and return to the first page (figure 8).

Now you need to click on step 4 “select content”.

You will see an image of the first page with a rectangle around the text; this is the rectangle that ScanTailor automatically selected according to its algorithms (figure 9). You will be able to see right away whether ScanTailor was correct. Maybe on some pages text will be visibly cut off, or not included in the rectangle. In order to correct all this, you will now flip through all the pages in your project and correct all such possible errors. You will also be able to immediately see and correct problems created at any previous steps (1-4), such as incorrect splitting of double pages.

In the page shown in figure 9, everything is okay, so you go to the next page. To flip to the next page, press PageDown or “W” on the keyboard. To go to the previous page, press PageUp or “Q” on the keyboard. (Or you can use the mouse wheel in the right column with thumbnails and then click on the thumbnails.)

Note the long horizontal button over the thumbnails; this is the “scroll lock” button. If this button is pressed, the thumbnail column will always show the page you are currently working on. Otherwise you can scroll away from your currently active page, to look at some other thumbnails.

As you go through the pages or switch between different steps, you may have to wait a little bit as the display updates. Eventually, as you go through all the pages, you will probably find a page where there is some problem after the draft run. There are five main types of problems to be corrected; most frequently:

1. the content rectangle needs adjusting (some text is outside, or the rectangle is too big and includes some “noise”)
2. the page alignment needs adjusting (usually at the beginning or at the end of a chapter, when most of the text is at the top of page or at the bottom of page)
3. incorrect splitting (this may happen when the page contains complicated tables and so was split when it shouldn’t have been)
4. incorrect deskewing (usually this happens when the page contains no text but only some large shapeless illustration).
5. the scan was done incorrectly (e.g. the page was not completely scanned)

Let us see how these problems can be corrected.

4.4.1 Adjusting the content rectangle

You can see in figure 10 that the content rectangle is too small; some part of the text was not included. Drag the content rectangle by the mouse, until it is correct. The content rectangle, as a rule, should not include any white margins — the content rectangle should fit snugly around the text on the page. White margins will be added later automatically.

An exception to this rule is when the content is neither centered nor flushed on the page: e.g. a dedication appearing somewhere towards the upper part of the page. In this case, it is easiest to make the content rectangle a little bigger, so that the default page alignment (which is "flush top and center horizontally") produces good results.

Note: if some page in the book is completely empty, you should make sure that it has *no* content rectangle at all. If any content rectangle was selected, you should right-click on the page and click "remove content rectangle". This will speed up processing (a perfectly white page with correct sizes will be generated).

Tip: While going through page contents, you will have to switch frequently between the windows "select content" and "page layout". It is possible to switch between them by keyboard shortcuts: press "P" for "Page layout" and press "S" *twice* for "Select content". (The first "S" will get you to "Split pages".)

4.4.2 Adjusting the page alignment

The page alignment options (see figure 7) are first, the sizes of the margins, and second, the alignment of the content rectangle. The default options are fine for most cases. (Note that the page will not be aligned unless the check is set on the "align with other pages" checkbox.)

Sometimes the beginning of a chapter has text that is flushed down on the page. You will have to correct the page alignment manually to "flush down" for these pages, while for most other pages, most probably, "flush up" is the right alignment option.

Sometimes you have a page with only very little text, or text that is only at the bottom of the page, or only at the top. You need to adjust the alignment of the page, or adjust the content rectangle so that it is aligned properly. For example, see figure 11.

In this case the default page alignment (which is "flush to top, center horizontally") will produce undesirable results (see figure 11, right). You can make this page "centered", but this is also not quite what you need. The easiest is to adjust the content rectangle so that it is larger and is aligned properly when flushed up and centered horizontally. Then you click on "page layout" and see something like figure 12 right.

4.4.3 Adjusting the page sizes

If you notice in the “page layout” preview that the final image has very wide or tall white margins on all pages, it means that some page is too wide or too tall, and ScanTailor has adjusted all pages so that they have the same pixel sizes. This can be fixed in two ways: first, you can exclude some pages from alignment by unclicking “align with other pages”. Second, you can make the margins smaller (say 0 mm) if you know that a significant amount of margin will be added by alignment anyway.

4.4.4 Adjusting the splitting

You can see in figure 13 that the page image does not contain a small part of the text. This cannot be fixed by adjusting the content rectangle because some text is entirely missing from the image. (Note that the content rectangle is also not quite right: it should be made a little taller to include the bottom part of the table frame. But here we discuss the problem that cannot be corrected by changing the content rectangle at this point.)

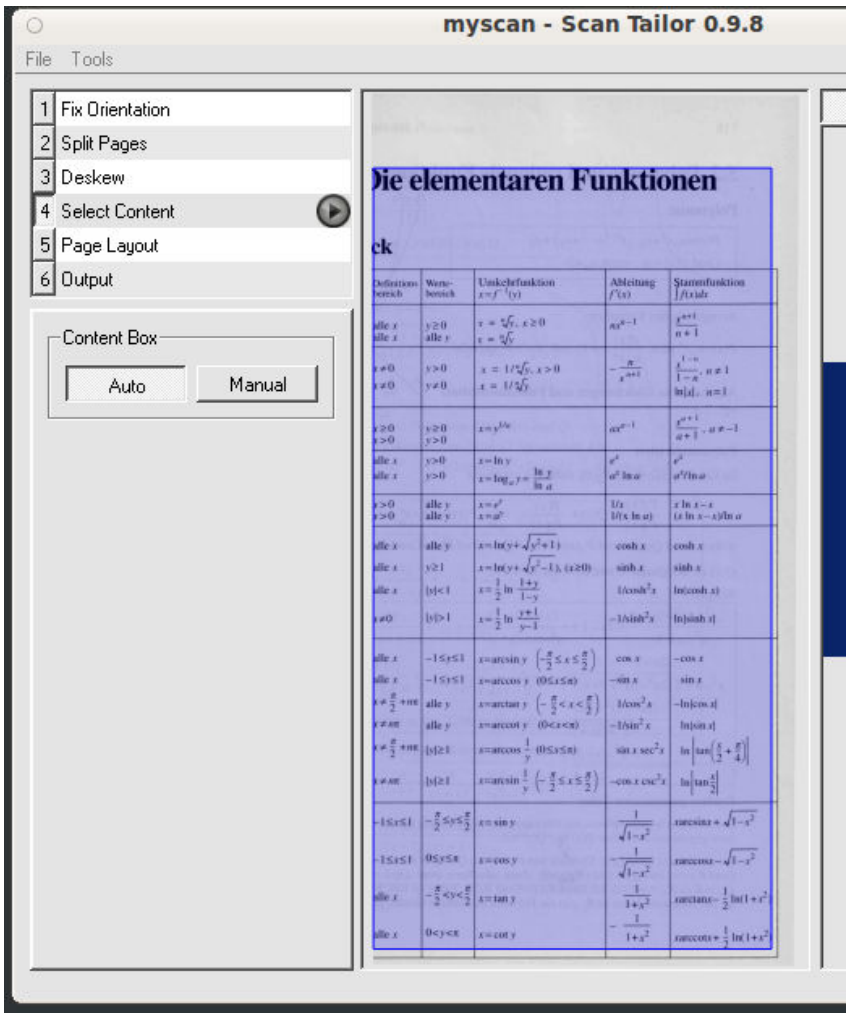


Figure 13: The left part of the page is missing and cannot be included in the content rectangle at all.

The problem is that some part of the text was cut away at the “splitting” step! Click on “split pages” and you will see something like figure 14.

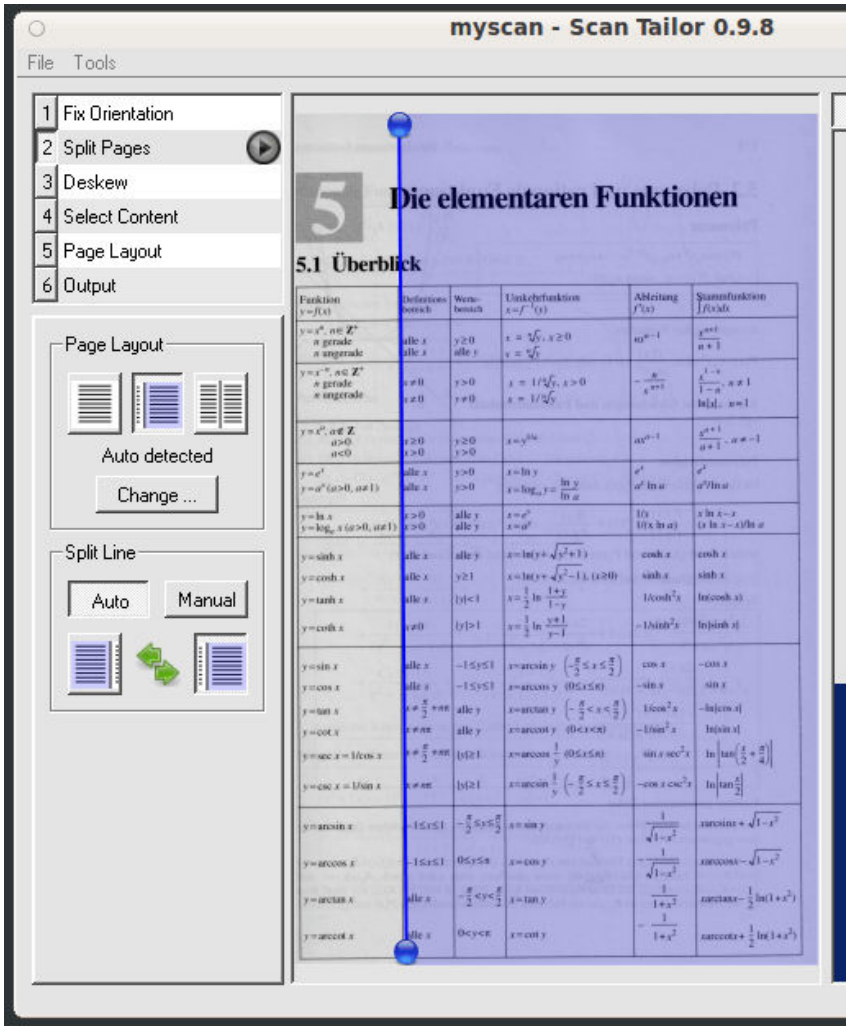


Figure 14: The “splitting” step shows the line of splitting. It was obviously incorrect.

Clearly, you need to drag the line of splitting to the left. After dragging that line, click again on “select content”. Now you will see a better content rectangle; still it needs to be adjusted a little, until you see something like figure 15.

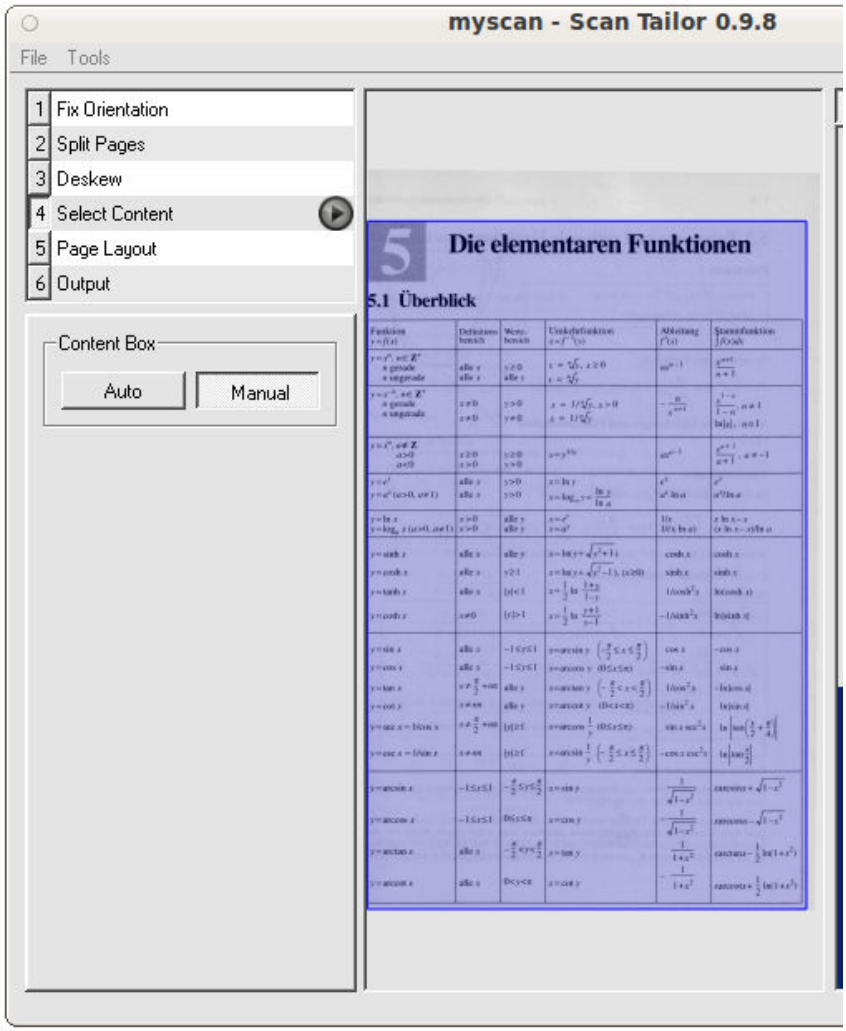


Figure 15: Problem corrected.

Note that in this example the scan did not actually need to be split. If this is the case with your scans, you can disable splitting entirely. To disable splitting, go to “Split pages” and select “Change...”, then “Mode: Manual” and “Scope: All pages” to apply this setting to all pages.

4.4.5 Adjusting the deskewing

This is a rare problem. If you see that the page image is still significantly skewed at the “select content” step, you need to click on the “deskew” step and drag the blue anchor point with the mouse until the page angle is better. Then you have to click again on “select content” and adjust the rectangle if necessary.

4.4.6 Replacing scans in the project

Finally, you might discover that you scanned some pages incorrectly (e.g. some part of the page was off the scanner glass). Then you can rescan that page

and add the new TIFF file to the project. Right-click on the thumbnail of some page; you will see a menu “Insert before”, “Insert after”, “Remove”. This allows you to remove incorrect scans and insert new, corrected scanned pages into the project (although this is done one page by one page, so if you want to add a lot of pages, it is better to start a new project).

Notes about removing or adding scans:

- When you remove pages from the project, the scans are not actually removed from the disk. Also, you can remove only one page from a split double-page scan, if necessary.
- It is advisable *not to remove any empty pages* in the middle of the book, because removing these pages will break the numeration of the pages. Empty pages will take practically no space in the final file. However, it is better to remove empty pages at the very beginning and at the very end of the book.
- When you add pages to the project, the new pages will have no processing steps already run on them, while other pages might be already partially processed. So, for instance, the new pages will appear to have the default page layout settings, and you will have to run all the steps on them, including “select content” and “page layout”.
- When a page has been removed from project, and this page was part of a two-page sheet, the settings for the other half of the sheet will be lost! You will have to click on that page and run again the “content selection” and “page layout” steps.

4.5 Final run and final check-up

After going through all the pages and correcting the layout errors, you need to return to the first page and click on the last step, “output”. You will see, after a somewhat longer waiting, a final version of the first page and the output options (see figure 16). The best options are: 600dpi, black/white mode, and slight despeckling (small “broom”), as is the default.

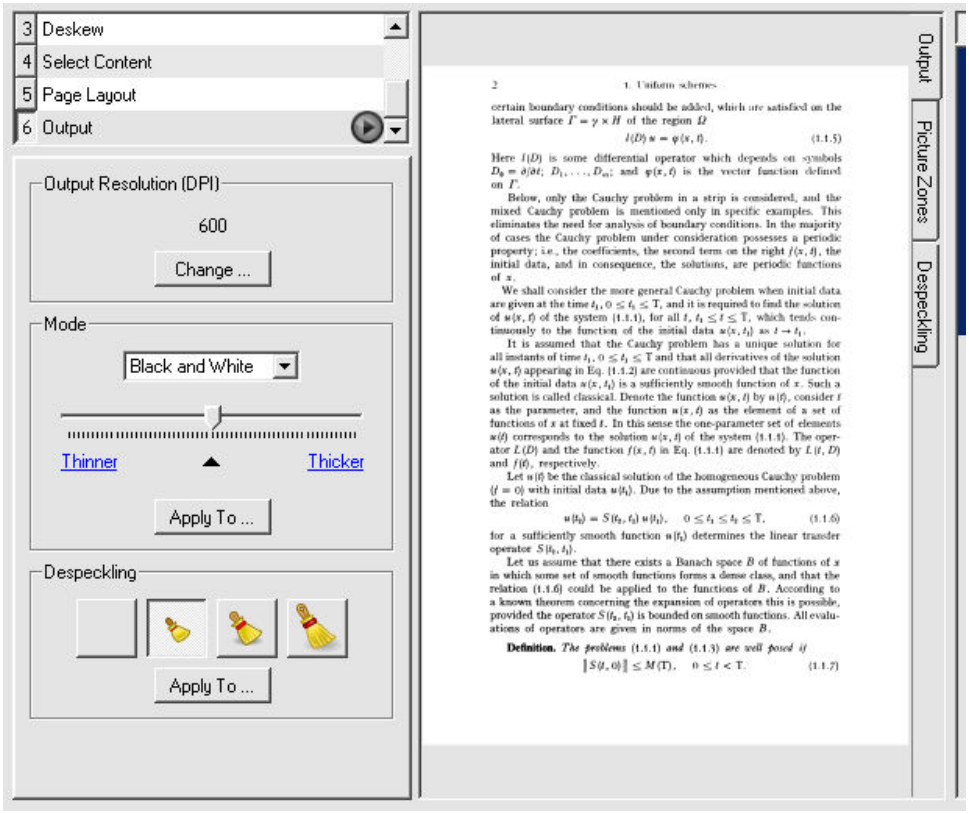


Figure 16: Output options at the last step.

You should check if the brightness of the final image is okay. If you see that the final picture is too dark and has lots of black dots around the text, you should move the slider towards the “thinner” setting and wait to see the new page image. If some of your scans are darker than others, you should scroll to them and click on their thumbnail; this will prepare the final image and you can then check whether it is too dark. Same if your final images are too light.

Note: If you see that some options are incorrect (e.g. Black/white should be changed to Mixed or Color) while ScanTailor is still calculating the image, you should change the options without waiting. Changing an option while a process is running will automatically restart the process for this page!

If your book has important greyscale or color illustrations on some pages, see section 4.6 for information about processing those pages. The other pages, which contain only black/white material, will still have to be processed as I will now describe.

If your book has all black/white text or black/white diagrams and no greyscale or color illustrations, you just need to adjust the brightness so that the image is sufficiently sharp and not too thin, not too thick.

You can view the page images at higher zoom: zooming is done by mouse scrolling. Scroll up for zooming in, scroll down for zooming out. Then you can see how the letters actually will look in the final image. You should expect smooth letter shapes at this point (ScanTailor applies some smoothing algorithms to the scans).

Note that you can apply the brightness settings to all pages at once, or only to selected pages (select them with the mouse in the thumbnail column at right), or only to pages after the current one. It is important to remember that the settings you click on the "output" window (or anywhere else in ScanTailor) are *only for the current page* unless you press "Apply" and select "Apply to all pages" or something else.

You are basically almost done at this point! Click again on the first page thumbnail, so that you see the first page, and then click on the "play" button to the right of "output". This will start the automatic processing of all pages. This operation is the "final run", which may take an hour or more (maybe about 15 seconds per page).

After this operation is done, you can do a final check-up of the pages. If the images for some pages are somehow still not correct, you can go back to any step and re-do it.

If your pages are all black/white, the only possible problems are these:

- Final image is too thin/too thick on some pages where the brightness was for some reason different from that of all other pages.
- Despeckling has removed some dots that are actually part of the text.

You can flip through the pages while viewing the despeckling results: click on the "despeckling" tab in the output window. The red dots will show where ScanTailor removed dots from the image. If you see that ScanTailor removed dots that are not "dirt" but actually are points in the text, such as ". . ." somewhere, you should use a different "despeckling broom" or disable despeckling altogether (or make the image "thicker"). Usually, ScanTailor will be careful with despeckling, but there are some cases when despeckling needs to be disabled for some or all pages.

Note: it is advisable to save your project often while you are working on it. ScanTailor is a stable program, but Windows is not, so if your computer crashes for any reason, you will be able to continue right where you last saved.

When you are done, the final images are in the output directory as a bunch of TIFF files. These files will be in 600dpi and black/white, so they will be much smaller than your original greyscale scans. This concludes the processing of scans; the next step would be converting these scans to DJVU, see section 5.

4.6 Working with picture zones

Now let us see what you need to do if your book *does* have some greyscale or color illustrations.

If your book has *a lot* of colored text (e.g., all chapter headings and all column titles on each page are in blue), you should consider *not* making them colored but reducing *all text* to black/white. The colored chapter headings are

not particularly useful; making them all black/white will not significantly decrease the usefulness of the book, but it will significantly decrease the amount of work you will have to expend on the file, and the final file size will be maybe half the size.

If there are some pages with important illustrations, you need to navigate to these pages and click on “output”. You will get to these pages if you flip through all the final images after the output run. Do not wait until the final image is produced and immediately click on “Mixed” in the “Mode” box.

In the “mixed” mode, ScanTailor will try to detect *automatically* where the greyscale or color illustration is located on the page. As an example, see figure 17.

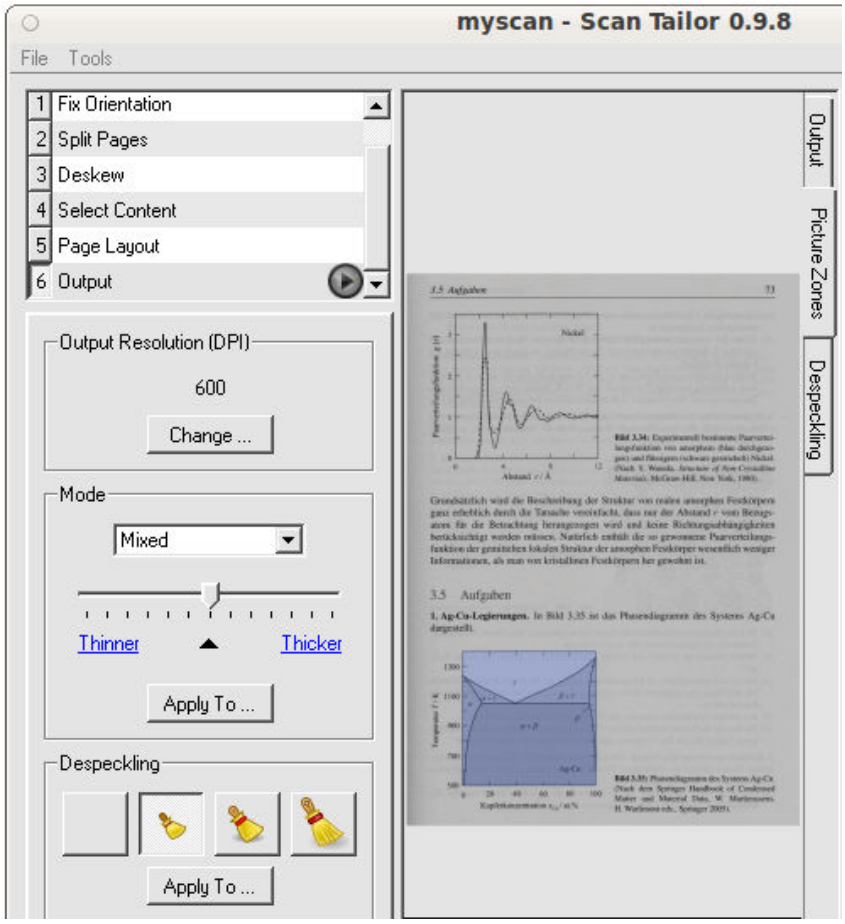


Figure 17: In the “mixed” mode, the illustration below is automatically detected as the “picture zone” and is shown to you in changing color when you click on the “Picture Zones” tab. Note that the upper illustration is purely black/white and was not selected as a “picture zone”.

You can also adjust the brightness of the final image in the “mixed” mode.

Sometimes ScanTailor guesses the picture zones somewhat incorrectly. Then you can draw your own picture zones with the mouse.

A few words about editing the picture zones. You can *add* new picture zones with boundaries made of straight lines. You *cannot* delete the automatically

found picture zone. But you can "*substract*" a picture zone from the zones already present. To do that, right-click on some point inside the picture zone and select "properties". Then you can select "subtract from all layers" or "subtract from the auto-layer".

If the automatically selected picture zone is very irregularly shaped, and if this is not right, perhaps the easiest thing to do is to draw a big picture zone around the automatically selected zone and select "subtract from auto-layer", so that the automatic picture zone is effectively removed, and then to draw your own picture zones and select "add to auto-layer". What you "add to auto-layer" will take precedence over what you "subtract from auto-layer". If you click "subtract from all layers", this is the highest "layer" and will subtract also from your added layers.

The other possibility is not to tinker with picture zones but encode *everything* as color. (The "color" mode.) In that mode, it is advisable to *check* the boxes "white margins" and "adjust luminosity". If you use this mode, the entire image will be saved as a picture zone. This will result in larger files, but is entirely acceptable and perhaps necessary if you have very complicated graphics that are not greyscale. Experiment and see what works best for your scans.

In any case, you can immediately see what the output will be for each given page. You will have to experiment until you find the right options. You can then apply these options at once to a group of pages or to all pages, by selecting the pages in the thumbnail column and pressing "Apply To" and then "To selected pages".

5 Encoding scans into DJVU

Once the processing of raw scans is finished, you have in the output folder a bunch of TIFF files which are (almost all) black/white at 600dpi. These TIFF files will take typically between 50 and 200 KB per page instead of about 4 MB that greyscale files took. By now you should have checked these TIFF files and made sure that the quality of the black/white images is good: the letters are sharp, have smooth shapes, there is little or no "dirt" etc. To check all that, you can view the TIFF files in a picture viewer (such as IrfanView) at high zoom.

Still, 50 to 200 KB per page is far too much. The next step is to encode these images to DJVU format; this will reduce their size dramatically, typically to 5-10 KB per page.

To make a good, well-optimized DJVU file, you need one of the two programs: either DjvuSolo version 3.1 or Djvu Document Express (DDE) 4.x, 5.x, 6.x or Djvu Document Express Enterprise (DEE) version 5.1 4.x, 5.x, 6.x.[12] The DDE and DEE programs are much faster than DjvuSolo, and DEE 5.1 can

[12]There is also a free software package called "djvulibre," but it cannot produce sufficiently well compressed DJVU files.

be configured to run in batch mode. On the other hand, DjvuSolo is a small and freely downloadable program that requires no setup. The results in terms of DJVU file quality from DjvuSolo and from DDE/DEE are pretty much the same if you set the options correctly.

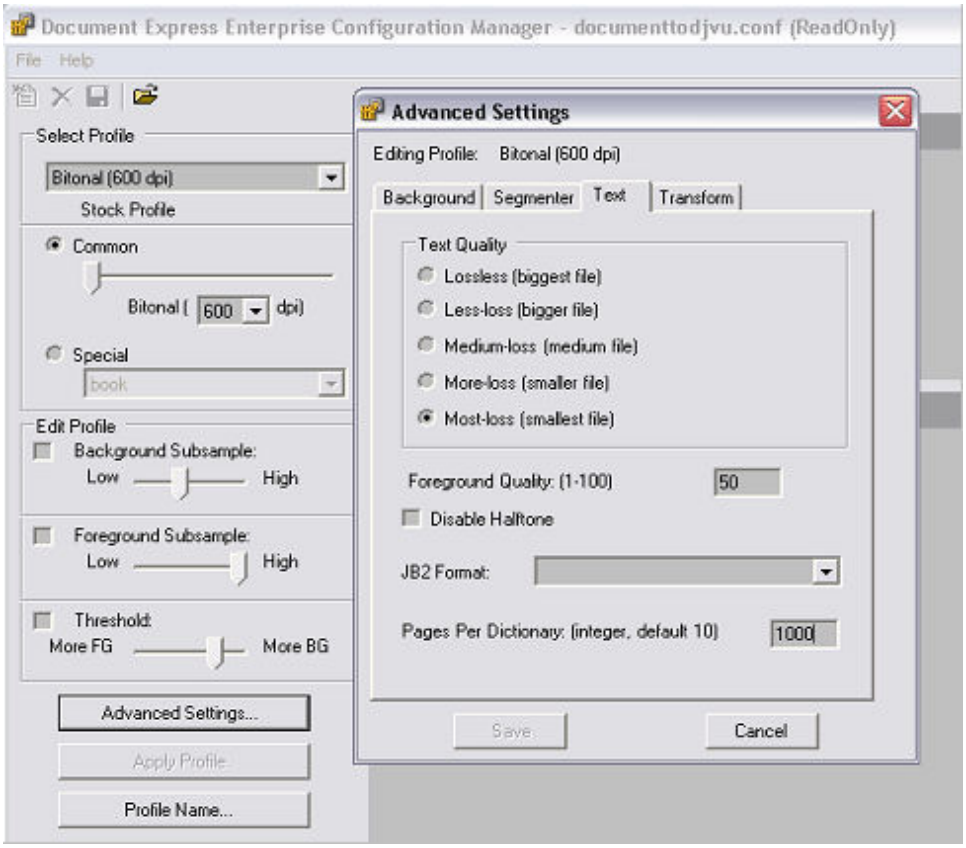
There are two ways of making DJVU files: one is by hand, another by batch. To make a DJVU file by hand, run DjvuSolo or DDE and click File→Open to open the first TIFF file. Then click Edit→Insert pages... and select all the other TIFF files. Please note: a selection box may have a bug in that you select many files by holding the **Shift** key and the mouse but they will be selected in the *inverse order* in the box. (This is a bug in a Windows dialog box.) Look at the text in the file name field and check that you are selecting the files in the correct order!

After "inserting the pages" you need to "Save as"... and select the "Bundled" format for DJVU and "Bitonal" option at 600dpi. You can also edit the file documenttodjvu.conf in the profiles directory and set pages-per-dict=100 or 200. The more pages per dictionary, the slower is the compression process, but the smaller the resulting file size.

Note that the "Bitonal" option (or "profile") in the DJVU encoders is intended for purely black/white scans, while "Scanned" option is intended for scans that have some (not many) colors but no photographs. Use the "Photo" option for photographs.

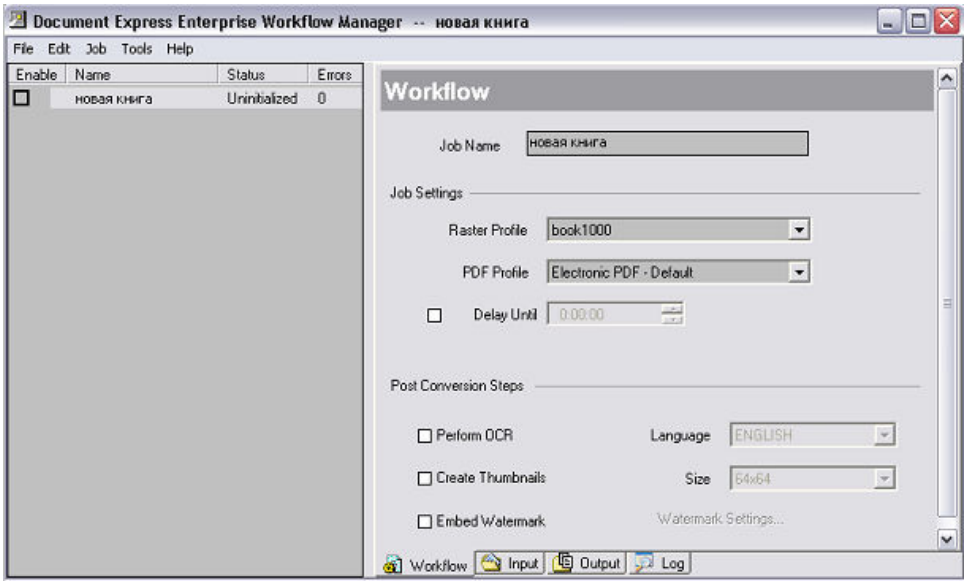
To make a DJVU file by batch, you need DEE 5.1.[13] First you need to create a special set of options (or "custom profile") for the DJVU encoding job. Run the Document Express Configuration Manager, choose the profile "Bitonal (600dpi)" from the list of profiles, click "Advanced settings", and you will see the following dialog.

[13]This is a rather large package; there exists a stripped-down version that takes only about 20MB on the hard disk.



Now choose the “Text” tab as shown above. In that tab, set “Pages per dictionary = 1000” (if this consumes too much RAM on your computer, or if this is too slow, set to 200 or 300 instead of 1000). Save the custom profile under a new name, say Bitonal-1. Do the same for the “Scanned (600dpi)” profile if you need to encode books with color drawings.

Now run the Document Express Workflow Manager. Load all the TIFF pages into it. In the “Job name” field, write the name of the book if you want. Choose the previously created custom profile in the list “Raster profile”.



Then click to the "Output" tab (the tabs are at the bottom of the window). In the list "Separate document(s)" choose "One document only." Tick the box under "Enable" at far left. Wait until the encoding is finished. You can also look at the "Log" tab to watch the progress. That's all; the DJVU file is created.

Do not delete the TIFF files yet! You may need to encode again if the DJVU file has some error. Also, the TIFF files are useful for OCR purposes (see section 6).

The result of DJVU encoding is a multipage DJVU file containing the entire e-book. You should rename that file to something sensible; not just math1.djvu. At the very least, the file name should contain the author's name, the title of the book, the publication year, and/or the ISBN number if available. This is just a little work, but it will be so much easier to share that file on the Internet if its name is sensibly chosen.

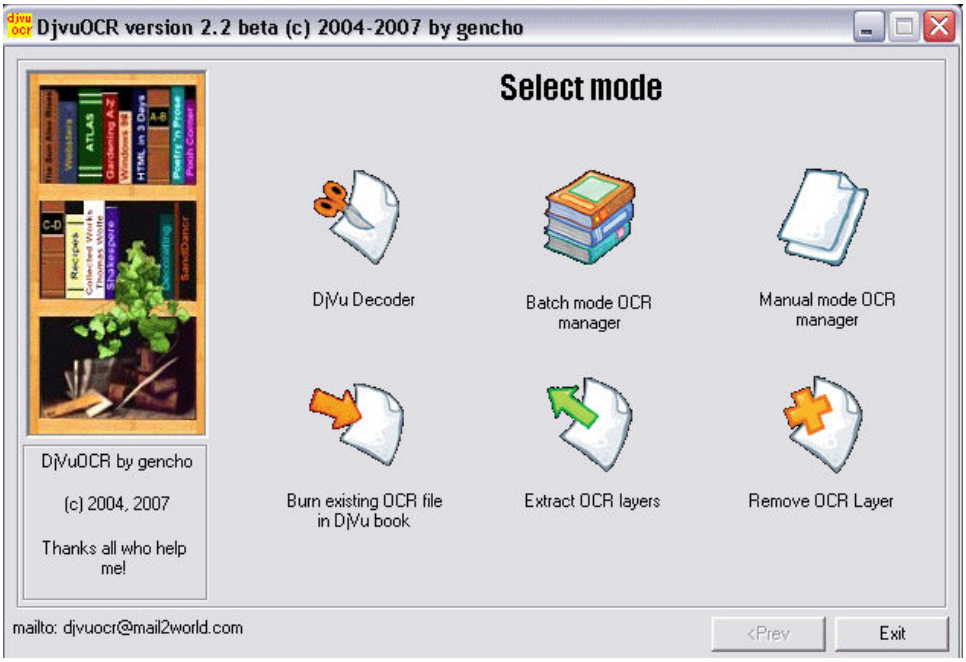
6 Creating text layer with OCR

Compared with the trouble needed to scan and process the book into a DJVU file, it is really peanuts to add OCR for it. An e-book with search is a lot easier to use.

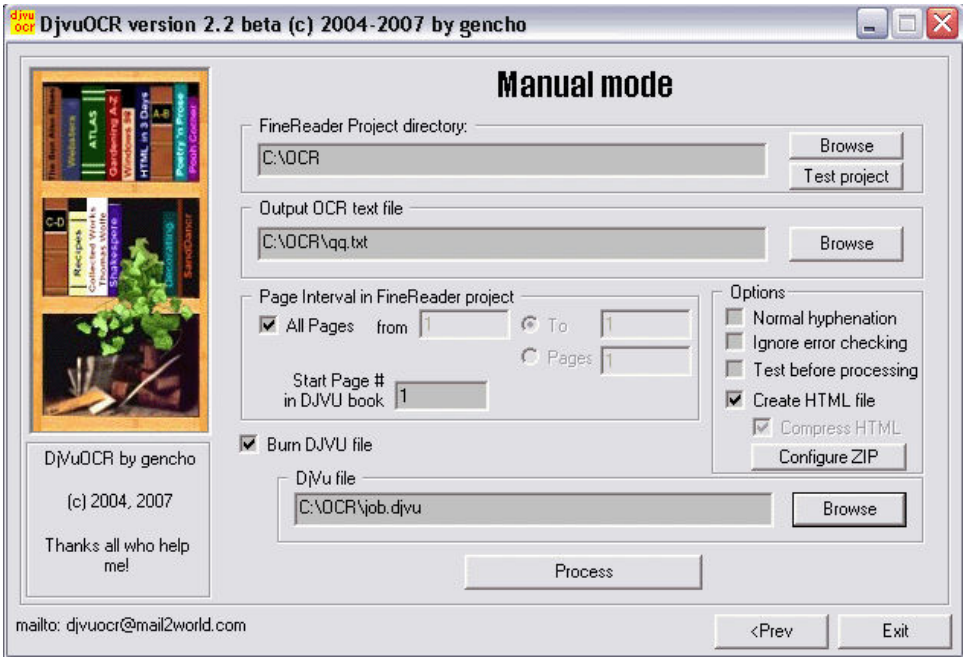
The search in DJVU files works only if the DJVU file has the so-called OCR layer. This layer is basically just a list of words stored inside the DJVU file in compressed form. You can create the OCR layer using two programs: FineReader and DjvuOCR. You need FineReader version 7 or 8.[14] It is okay to use even a trial or unregistered or evaluation version that you can download for free. The result of running FineReader will be a set of FineReader batch files. The wonderful program DjvuOCR created by Gencho will read these files directly, extract the OCR information, and insert it into DJVU files.

Suppose you have already created the DJVU file out of some TIFF files. Hopefully, you didn't delete the TIFF files. Load the TIFF files into a new batch in FineReader (keep in mind the problem with selecting many files at once!). Set the recognition language and press "Read all". When the OCR process is finished, click "Save batch". It is not recommended to edit the OCR text. Previous versions of DjvuOCR could not process FineReader batches if the OCR text was edited. The most recent version DjvuOCR 2.2, can deal with small edits. You should not rewrite large blocks of text; i.e. you should keep many original symbols in their original positions if you edit. Also you should not delete the end-of-line symbols, so that the number of lines in a paragraph remains the same. But we recommend that you do not edit the OCR text at all. After saving the FineReader batch, you can quit FineReader and run the program DjvuOCR.

[14] FineReader 9 is now available but it cannot add OCR to DJVU files, and there is no DjvuOCR support for FR 9.



This program has several functions; for example, "DjVu Decoder" will produce TIFF files out of DJVU in case you deleted your TIFF files, or if you are working with somebody else's DJVU file. For now, you will use only the "Manual mode OCR manager." Click that, and you get the following window.



Select the directory where the FineReader batch is located in the "FineReader Project directory" field. "Output OCR text file" will be the name of the new file; it doesn't matter what that name is. Tick the "Burn DJVU file" box and select the DJVU file below; it means that the OCR data will be inserted ("burned") into the DJVU file. Click "Process", wait a few minutes, and that's all. Now the DJVU file is full-text searchable!

7 Adding book covers and color plates

It is reasonably easy to add a simple book cover. Just scan the book cover in 300dpi color, or even in 200dpi. Slightly blur the image in a graphics editor. Encode into DJVU using the profile "Photo(300)" or "Scanned." The resulting 1-page DJVU file needs to be inserted at the beginning of the DJVU e-book after all the other processing is finished. Usually the book cover should not be larger than 20-30 KB. It is probably not necessary to spend a lot of effort on making a great-looking book cover. Consider that the people who will read your e-book will spend most of the time reading the text rather than looking at the cover.

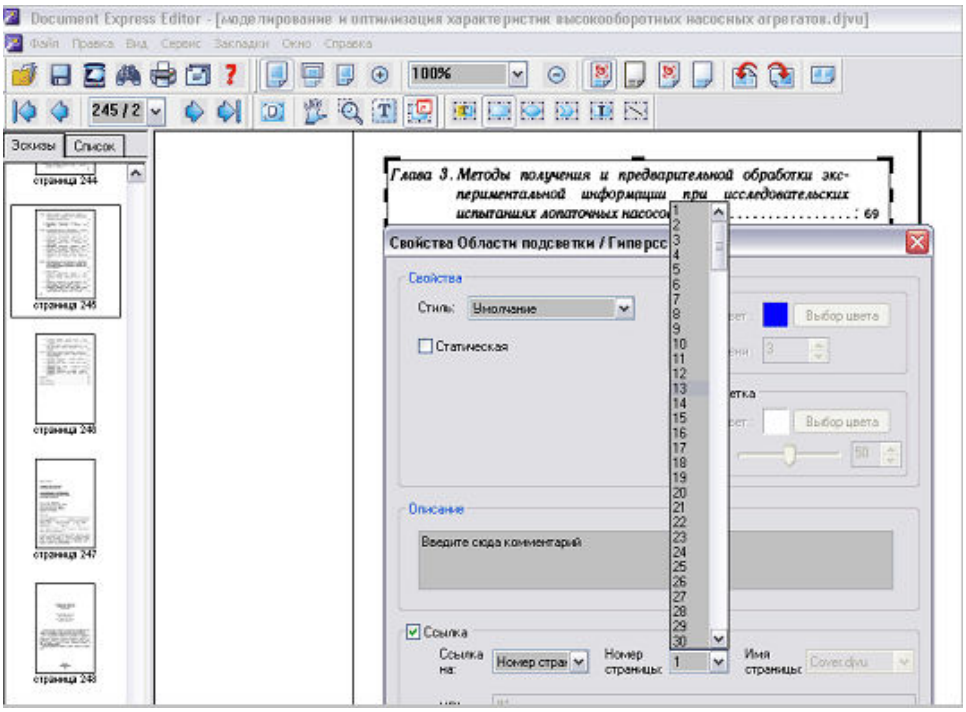
In the same way one can add *color plates*, that is, special pages that contain only color illustrations. Scan them separately and insert into the finished DJVU file after all other processing is done.

To insert or rearrange pages in a DJVU file, use DjvuSolo or DDE. Open the DJVU file, and you will see the thumbnails of the pages in the left column. You can simply drag the thumbnails to rearrange the pages; you can also "Cut," "Copy," and "Paste" pages or groups of selected pages, or delete pages. Use the menu Edit→Insert pages... to add more DJVU pages to an existing DJVU file. You can insert single-page or multipage DJVU files anywhere (before or after any page), as you need.

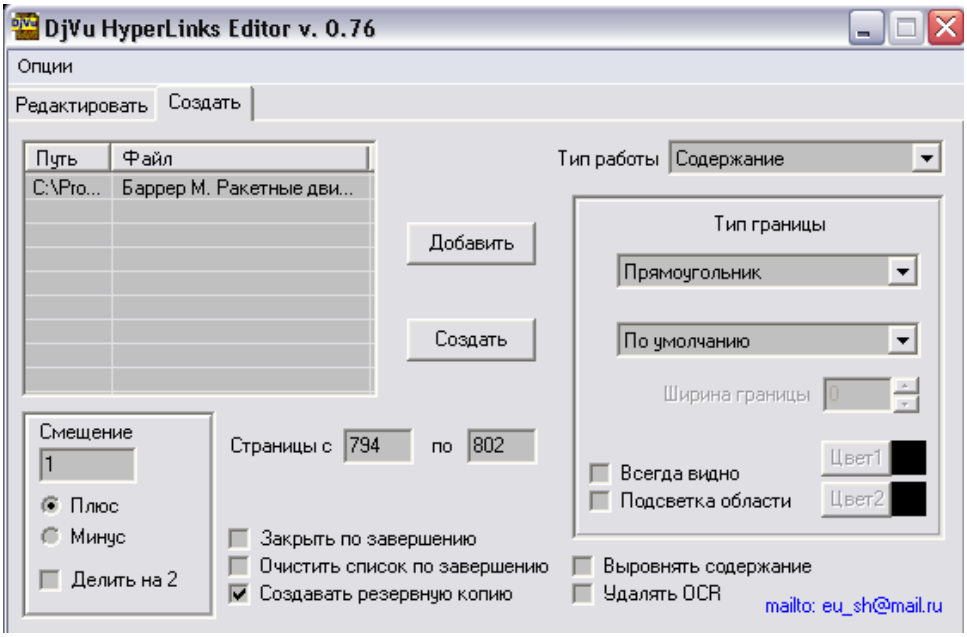
8 Adding hyperlinks and bookmarks

After finishing all the preceding work with the DJVU file (including OCR), you can add some hyperlink navigation to it. There are two ways of adding hyperlinks.

The first is to use the DjvuSolo or Djvu Editor programs and add hyperlinks by hand. Usually, one adds hyperlinks to pages in the table of contents for easier navigation. In DjvuSolo or Djvu Editor you can select any rectangular area on any page and then insert a hyperlink to a different page of the DJVU file. The user will go to this page when clicking anywhere in the area. Note that the hyperlink will point to a page number, so adding hyperlinks has to be done *after* any changes to the page order or *after* inserting any additional pages into the DJVU file. So if you want you can sit and make some rectangular areas into hyperlinks until you are blue in the face.



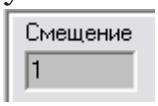
The second way to add hyperlinks is semi-automatic, using the program DJVU Hyperlinks Editor.[15] Run the program and you will see the following window.

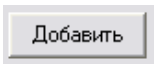


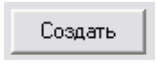
First you need to specify options for the hyperlinks Then you need to specify the page range (

) in which the table of contents is located in the DJVU file. These are DJVU page numbers, which may be different from the page numbers printed in the book and in the table of contents (e.g. because there are some pages taken by the cover and by the front matter). To compensate for this, usually one needs to add a certain offset to the page number; for

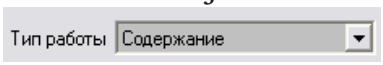
[15] This program has only the Russian-language interface.

instance, page 10 in the printed book may be actually page 11 in the DJVU file because one page is taken by the cover.[16] Then you need to enter the corresponding offset into the box 

("offset"). Now that all options are enterd, press the button 

(which means "Add"). This will add a new DJVU file to the list in the left panel; the current options will apply to that file. You can now set different options and add a different file. Finally, press the button 

("create"). This will insert the hyperlink information into all the DJVU files.

Similarly, one can create hyperlinks in the subject index. One needs to select a different entry in the drop box 

. The default entry as shown means "Table of contents." Other entries mean that you want to process the subject index. The same settings apply.

After finishing the processing, one should view the DJVU file and check that the hyperlinks were added correctly. The program relies on the OCR text for determining the page numbers for hyperlinks. So any errors in OCR may lead to errors in the position or targeting of the hyperlinks.

A Where to download software

Name of program	*Download site*	*Status*
IrfanView 4.1	www.irfanview.com	free
ScanTailor 0.9.8	scantailor.sf.net	free
ScanKromsator 5.9	www.djvu-soft.narod.ru	free
DjvuSolo 3.1	www.djvu-soft.narod.ru	free
Djvu Editor 4.x, 5.x, 6.x (DDE/DEE)	www.djvu-soft.narod.ru	nonfree
FineReader 7.x, 8.x	www.abbyy.com	trial
DjvuOCR 2.2	djvuocr.ucoz.ru	free
Djvu Hyperlinks Editor	www.djvu-soft.narod.ru	free

Big thanks to **monday2000** for creating the website djvu-soft.narod.ru

Note for Linux users: All the programs in this table work reasonably well under the standard Windows emulator (wine). However, some programs (IrfanView, DDE/DEE, FineReader) may fail to install if you run "setup.exe" for those programs. You need to get "portable" or "installed" versions of these programs that do not require running an installer. ScanTailor has a native Linux version that can be compiled from the sources.

[16] This is the Russian convention where the page numbering starts right away from the first page of the book. In the Western typography the front matter usually has separate roman numbering, so typical offsets will be not 1 but between 10 and 20.

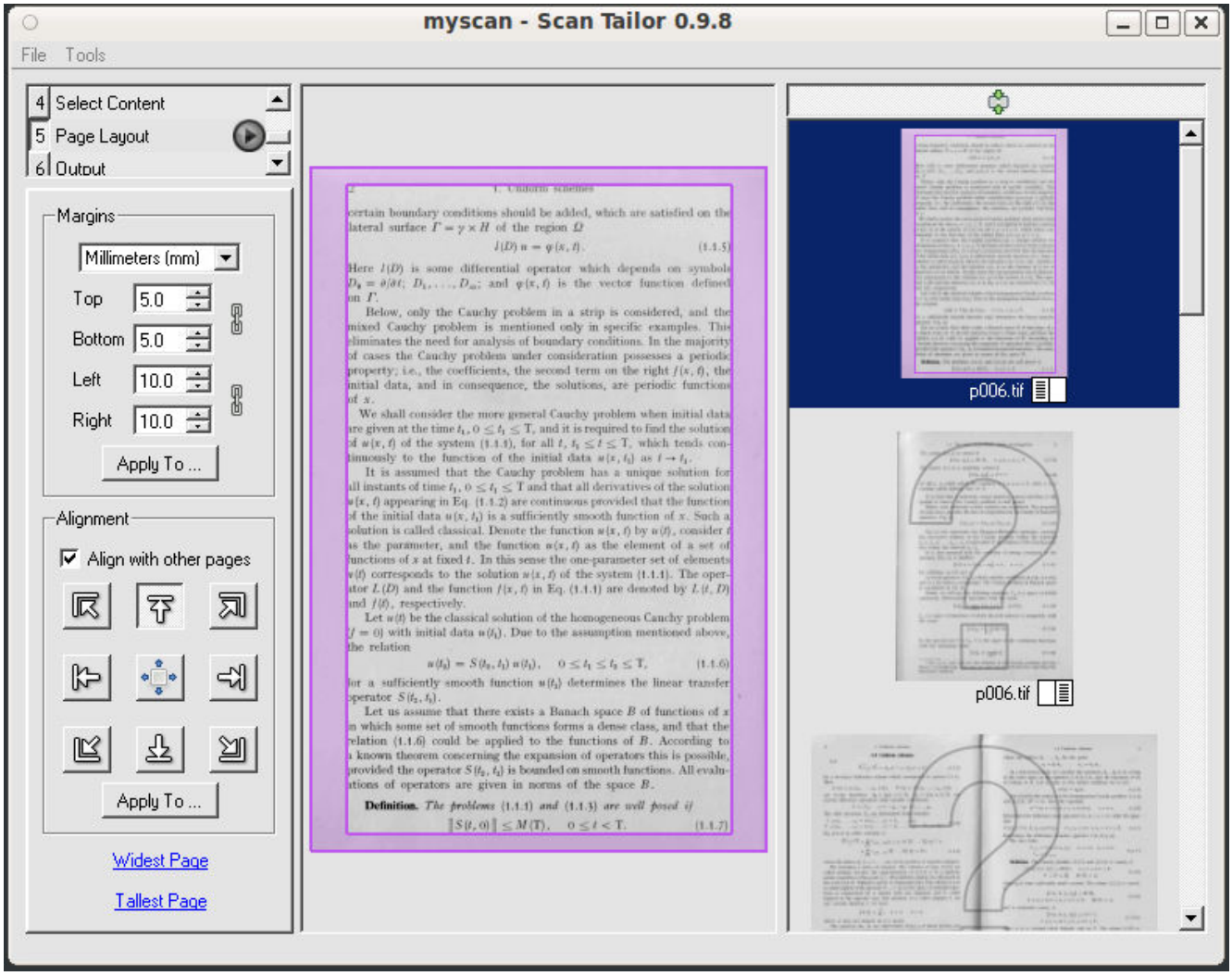


Figure 7: After clicking on "page layout" while on the first page. The big question marks on the thumbnails mean that these pages have not yet had this step ("page layout") performed on them. The "Alignment" symbols mean the centering or flush-centering of the page in various directions. Press on them to see immediately what effect these options would have on the final appearance of the page.

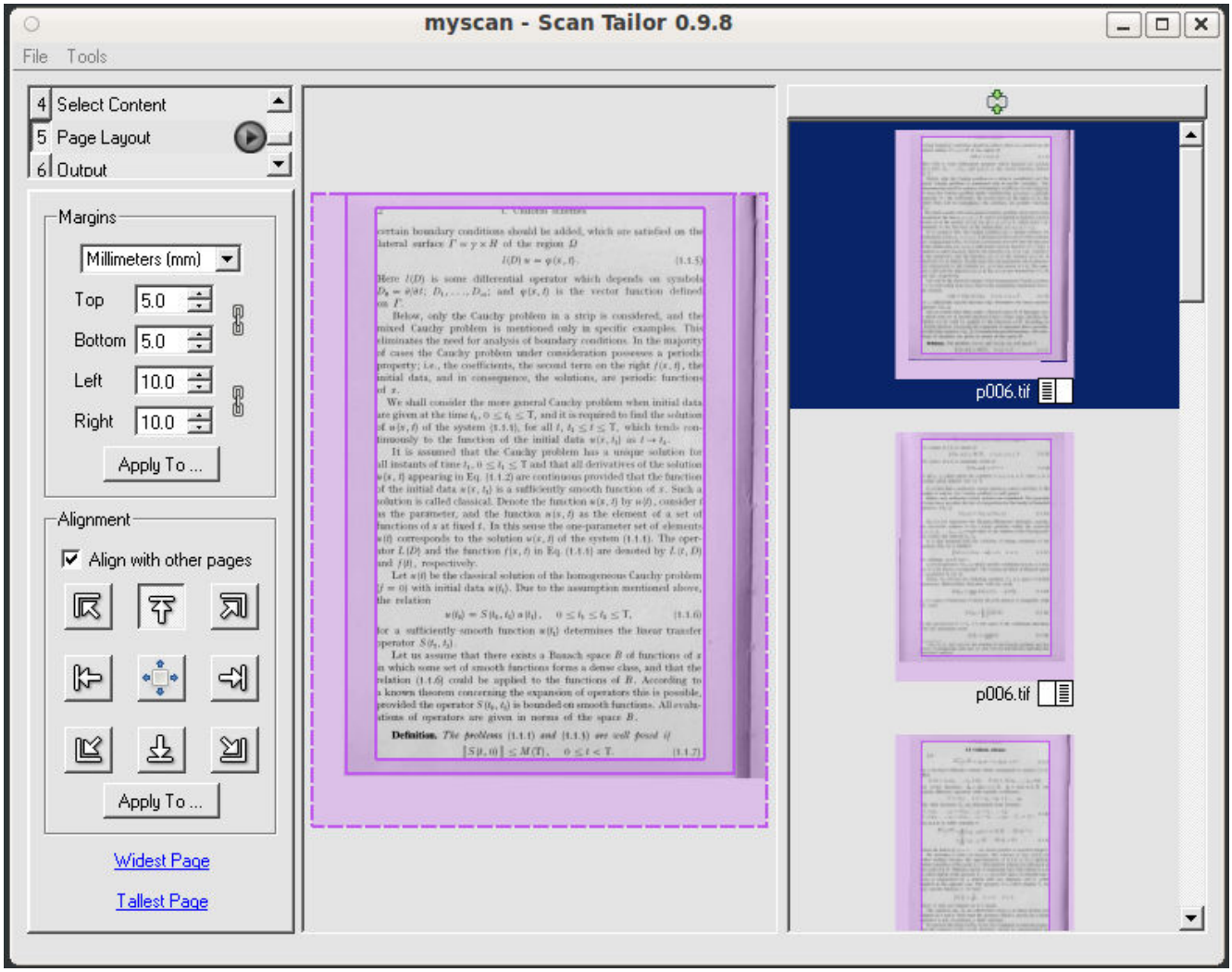


Figure 8: After the “draft run” you are again at the first page. The big question marks on the thumbnails are gone.

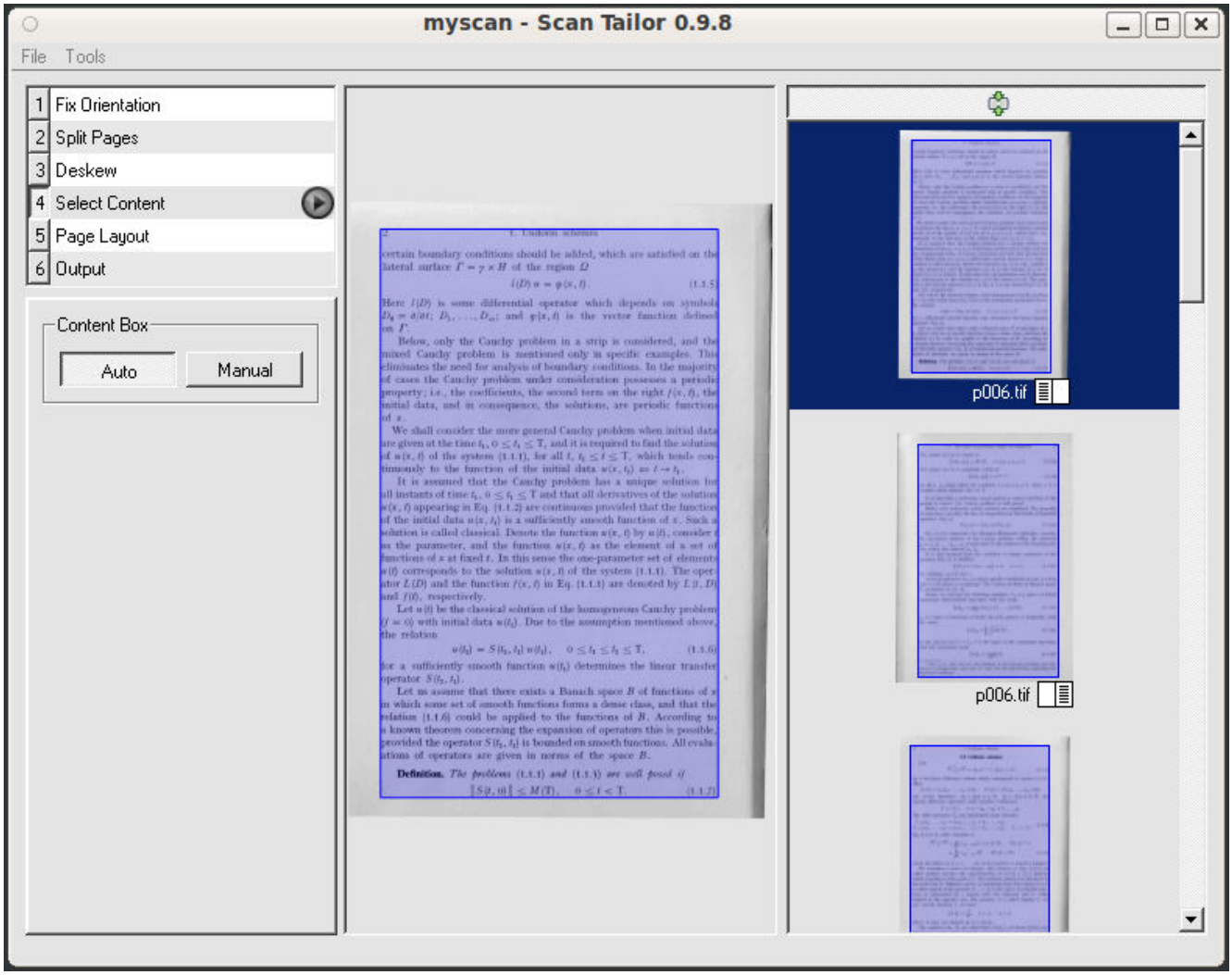


Figure 9: After you click on “select content” you can inspect the content rectangle. In most cases (like on this page), the content is detected perfectly.

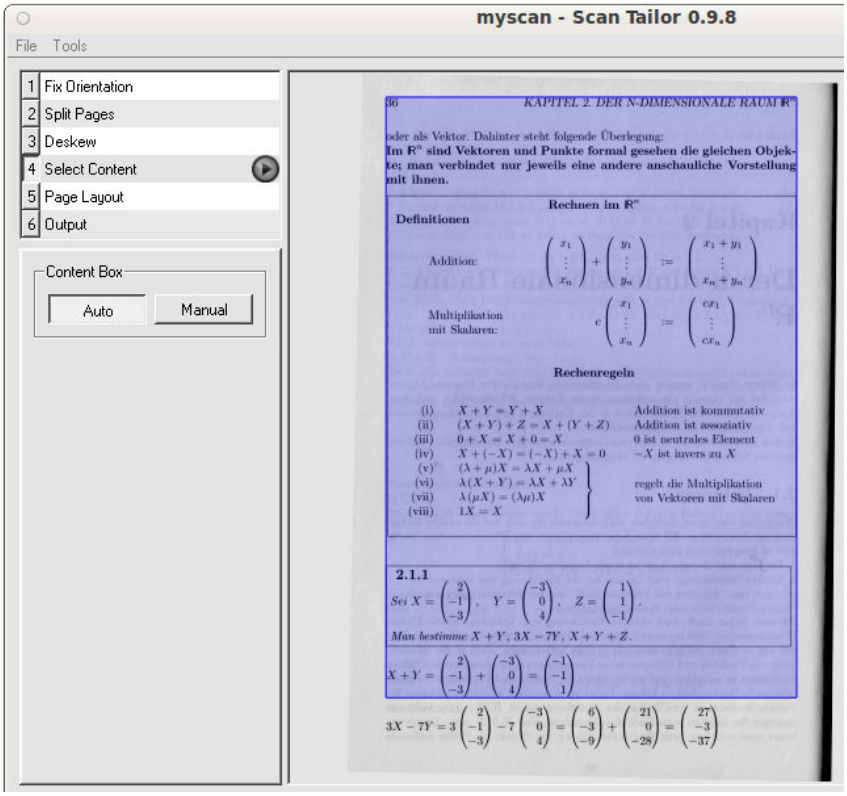


Figure 10: In this case the content rectangle is too small. You need to adjust it by dragging with the mouse.

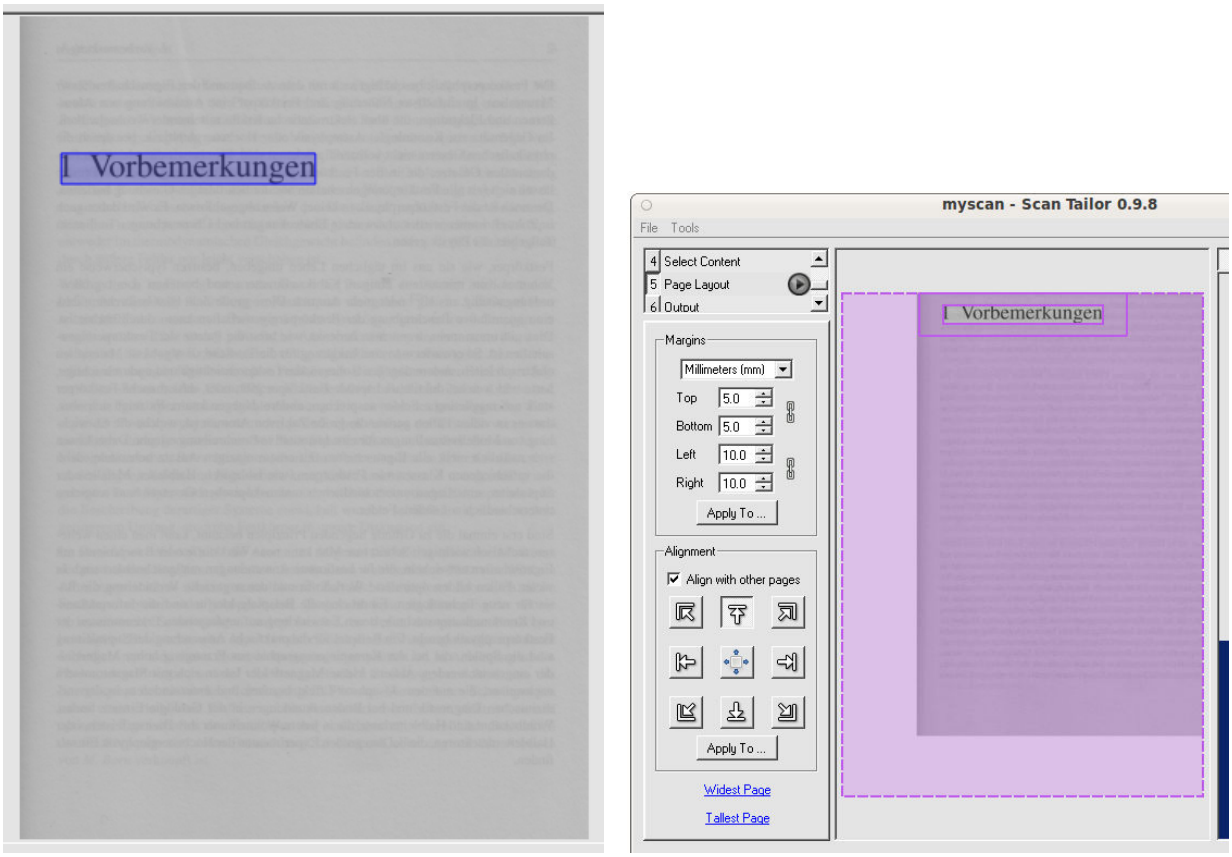


Figure 11: The content rectangle is correct but very small (see on left). The default page alignment will flush this rectangle to the top of the page and center it, which is not what is desired.

Figure 12: An enlarged content rectangle (left) produces good page layout (right) like in the original printed page.

Index

A3 scanner, 12

color plates, 38

deskewing, 11
DJVU, 4, 33
    dictionary, 34
    OCR layer, 36
    rearrange pages, 38

FineReader
    problems, 5

illustrations, 4
IrfanView, 7

JPEG, 8
    digital artifacts, 9
    problems, 8

kromsating, 14

quality, 3, 4

ScanKromsator, 5, 13
    cutters, 15
    draft run, 14
    main run, 18
    picture zones, 18
scanning, 11, 12
    disk space, 12
    greyscale, 4
    with digital camera, 5
ScanTailor, 5, 19

TIFF, 8

upsampling, 4, 18
using Linux, 40

VueScan, 9